Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 2 (2026) pp: 22168-22175
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analisis Perbandingan Metode *Clustering* Untuk Seleksi Penerima Bantuan Pendidikan Berbasis Data Mining

Fajrin[1], Apriani[2], Ria Rismayati[3]
[1]Program Studi Ilmu Komputer, Fakultas Teknik, Universitas Bumigora
[23]Fakultas Teknik Program Studi Ilmu Komputer, Universitas Bumigora
ifaja816@gmail.com*, apriani@universitasbumigora.ac.id, riris@universitasbumigora.ac.id

Abstrak

Penentuan penerima bantuan pendidikan yang masih dilakukan secara manual rentan menimbulkan subjektivitas dan ketidaktepatan sasaran akibat keterbatasan dalam mengolah data siswa secara menyeluruh. Penelitian ini bertujuan membandingkan kinerja lima metode clustering, yaitu K-Means, Fuzzy C-Means, Gaussian Mixture Model, Spectral Clustering, dan Agglomerative Hierarchical Clustering, dalam mengelompokkan calon penerima bantuan pendidikan berdasarkan kondisi sosial ekonomi, serta menentukan metode terbaik. Penelitian menggunakan pendekatan kuantitatif deskriptif komparatif dengan dua dataset, yaitu data siswa SMKN 1 Lambu sebanyak 1.241 data sebagai objek utama dan dataset Academic Performance Retention dari Kaggle sebanyak 574 data sebagai pembanding. Atribut yang digunakan meliputi jenis tempat tinggal, alat transportasi, jumlah saudara, pekerjaan dan pendapatan orang tua, serta jarak rumah ke sekolah. Tahapan penelitian meliputi pembersihan data, transformasi, normalisasi menggunakan RobustScaler, reduksi dimensi dengan Principal Component Analysis, penerapan lima metode clustering dengan tiga cluster, serta evaluasi menggunakan Silhouette Score dan Davies-Bouldin Index. Hasil menunjukkan Agglomerative Hierarchical Clustering memberikan performa terbaik pada dataset utama dengan Silhouette Score 0,652852 dan Davies-Bouldin Index 0,597686, mengungguli K-Means, Fuzzy C-Means, Gaussian Mixture Model, dan Spectral Clustering. Metode ini menghasilkan cluster yang lebih kompak dan terpisah antarkelompok, sehingga direkomendasikan sebagai dasar sistem pendukung keputusan penentuan penerima bantuan pendidikan yang lebih objektif dan tepat sasaran.

Kata Kunci: *Clustering, K-Means, Fuzzy C-Means, Gaussian Mixture Model, Hierarchical Clustering, Bantuan Pendidikan*

1. Latar Belakang

Pendidikan merupakan hak dasar setiap warga negara yang berperan penting dalam pembentukan sumber daya manusia berkualitas [1]. Untuk menjamin pemerataan akses tersebut, pemerintah dan berbagai lembaga menyediakan program bantuan pendidikan berupa beasiswa maupun subsidi biaya sekolah bagi peserta didik dari keluarga kurang mampu [2]. Keberhasilan program ini sangat bergantung pada ketepatan proses seleksi, karena kesalahan penentuan penerima dapat menimbulkan kecemburuan sosial sekaligus membuat bantuan tidak sampai pada siswa yang sebenarnya membutuhkan [3]. Pada kenyataannya, proses seleksi di banyak sekolah masih dilakukan secara manual dan hanya bersandar pada penilaian subjektif pihak sekolah, sehingga rawan *human error*, sulit menangani data dalam jumlah besar, serta menghasilkan keputusan yang tidak konsisten antarperiode [4]. Kondisi ini berdampak pada ketidaktepatan sasaran, karena indikator penting seperti kondisi sosial ekonomi keluarga dan akses transportasi tidak dianalisis secara menyeluruh [5].

Ketersediaan data siswa yang telah tersimpan secara digital di sekolah sesungguhnya membuka peluang bagi penerapan pendekatan berbasis data, namun tanpa metode analisis yang tepat data tersebut hanya berfungsi sebagai arsip administratif. Penerapan data mining pada berbagai instansi juga telah terbukti mampu mengubah data operasional menjadi informasi yang mendukung proses pengambilan keputusan secara lebih efektif [6]. Salah satu teknik *data mining* yang relevan untuk mengatasi persoalan ini adalah *clustering*, yang mengelompokkan data berdasarkan kemiripan karakteristik tanpa memerlukan label kelas sebelumnya. *K-Means* menjadi metode *clustering* yang paling banyak diterapkan dalam konteks pendidikan karena kesederhanaan algoritma dan kecepatan komputasinya, misalnya dalam pengelompokan siswa berdasarkan kondisi sosial ekonomi [7]. Meskipun demikian, sifat *hard clustering* pada *K-Means* membuatnya kurang fleksibel ketika data memiliki karakteristik yang saling tumpang tindih antarkelompok.

Kondisi tumpang tindih semacam ini banyak dijumpai pada data sosial ekonomi siswa, misalnya ketika dua keluarga memiliki tingkat pendapatan yang berdekatan namun berbeda pada aspek lain seperti jumlah tanggungan

atau jarak tempuh ke sekolah. Pada situasi tersebut, penetapan satu label cluster secara kaku berisiko menempatkan calon penerima bantuan pada kategori yang kurang mewakili kondisi sebenarnya, sehingga dibutuhkan pendekatan yang mampu mengakomodasi derajat keanggotaan yang tidak tegas antarkelompok.

Keterbatasan tersebut mendorong penggunaan *Fuzzy C-Means* (FCM) sebagai pendekatan *soft clustering* yang memungkinkan setiap data memiliki derajat keanggotaan pada lebih dari satu *cluster*, dan terbukti efektif pada data sosial ekonomi yang batas kategorinya tidak tegas [8]. Di sisi lain, *Gaussian Mixture Model* (GMM) menawarkan pendekatan probabilistik yang mampu menangkap struktur data kompleks dan tidak simetris [9]. Sayangnya, ketiga metode ini pada umumnya dikaji secara terpisah pada masing-masing penelitian, sehingga perbandingan langsung ketiganya dalam satu konteks penelitian penelitian mengenai pengelompokan penerima bantuan pendidikan juga menunjukkan bahwa pemilihan algoritma clustering berpengaruh terhadap kualitas hasil pengelompokan yang diperoleh [10], khususnya seleksi penerima bantuan pendidikan, masih sangat terbatas [7].

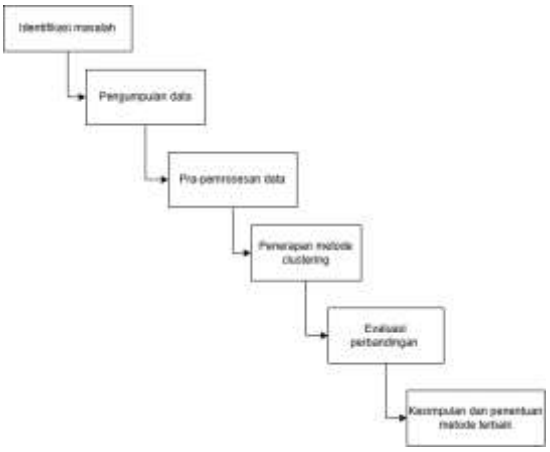
Ketimpangan kajian tersebut memperlihatkan bahwa evaluasi K-Means, FCM, dan GMM masih dilakukan pada konteks data yang berbeda-beda, sehingga belum dapat disimpulkan metode mana yang paling andal ketika diterapkan pada satu jenis data yang sama, khususnya data sosial ekonomi siswa dengan atribut campuran antara kategorikal dan numerik. Ketiadaan perbandingan langsung ini turut menyulitkan pihak sekolah dalam menentukan algoritma yang paling sesuai untuk diadopsi ke dalam sistem seleksi bantuan pendidikan mereka.

Selain ketiga metode di atas, *Spectral Clustering* yang memanfaatkan teori graf terbukti mampu meningkatkan kualitas pengelompokan pengguna pada sistem rekomendasi dibandingkan pendekatan tanpa *clustering* [11], sementara *Agglomerative Hierarchical Clustering* dilaporkan menghasilkan kualitas *cluster* yang sangat baik (*Silhouette Score* 0,79) pada pengelompokan prioritas penerima bantuan rumah [12]. Kedua temuan ini menunjukkan potensi penerapan kedua metode tersebut pada data sosial ekonomi siswa, namun penelitian yang membandingkan kelima metode — K-Means, FCM, GMM, *Spectral Clustering*, dan *Hierarchical Clustering* — sekaligus dalam satu kerangka evaluasi untuk kasus bantuan pendidikan belum ditemukan pada literatur yang ada.

Berdasarkan kesenjangan tersebut, penelitian ini menawarkan kebaruan berupa analisis komparatif langsung terhadap lima metode *clustering* pada dua dataset (data siswa sebagai objek utama dan dataset pembanding), dengan evaluasi konsisten menggunakan *Silhouette Score* dan *Davies-Bouldin Index* untuk menghasilkan rekomendasi metode yang paling sesuai sebagai dasar sistem pendukung keputusan penentuan penerima bantuan pendidikan. Penelitian ini bertujuan menerapkan dan membandingkan performa kelima metode tersebut serta menentukan metode dengan kualitas pengelompokan terbaik berdasarkan hasil evaluasi.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional di bidang *data mining*. Pendekatan ini dipilih karena penelitian berfokus pada pengolahan data numerik serta pengukuran performa metode *clustering* secara objektif menggunakan metrik evaluasi internal. Secara garis besar, penelitian dilaksanakan melalui tahapan yang saling berurutan, yaitu identifikasi masalah, pengumpulan data, pra-pemrosesan data, penerapan model *clustering*, evaluasi hasil pengelompokan, penarikan kesimpulan dan rekomendasi, hingga tahap penerapan hasil ke dalam sistem pendukung keputusan. Alur metode penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Metode Penelitian

DOI: https://doi.org/10.31004/riggs.v5i2.11894

2.1 Identifikasi Masalah

Tahap awal penelitian dimulai dengan melakukan identifikasi terhadap permasalahan yang terjadi dalam proses seleksi penerima bantuan pendidikan. Proses seleksi yang masih dilakukan secara manual dinilai kurang efisien dan berpotensi menimbulkan subjektivitas dalam pengambilan keputusan. Selain itu, belum adanya sistem berbasis data yang mampu mengelompokkan siswa secara otomatis menjadi salah satu kendala dalam menentukan prioritas penerima bantuan.

2.2 Pengumpulan Data

Data penelitian bersumber dari dua dataset. Dataset utama merupakan data siswa SMKN 1 Lambu sebanyak 1.241 data dengan 15 variabel, diperoleh melalui kegiatan magang penulis di sekolah tersebut. Dataset kedua adalah *Academic Performance Retention* sebanyak 574 data dengan 17 variabel yang diperoleh dari platform Kaggle dan digunakan sebagai data pembanding untuk menguji konsistensi kinerja metode *clustering* pada karakteristik data yang berbeda. Dari kedua dataset tersebut dilakukan seleksi atribut, sehingga hanya variabel yang berkaitan langsung dengan kondisi sosial ekonomi yang digunakan, yaitu jenis tempat tinggal, alat transportasi, jumlah saudara, pekerjaan orang tua, tingkat pendapatan orang tua, dan jarak rumah ke sekolah pada dataset utama, serta *residence location*, *residence type*, *parental education*, *parental income level*, *employment status*, dan *socioeconomic level* pada dataset pembanding. Atribut yang bersifat identitas pribadi tidak digunakan karena tidak memberikan kontribusi terhadap pembentukan pola *cluster*.

2.3 Pra-Pemrosesan Data

Tahap pra-pemrosesan data diawali dengan *data cleaning* untuk memeriksa nilai kosong dan data duplikat; hasil pemeriksaan menunjukkan kedua dataset tidak memiliki nilai kosong maupun duplikasi sehingga dapat langsung digunakan pada tahap berikutnya. Atribut kategorikal selanjutnya ditransformasikan menjadi bentuk numerik menggunakan teknik *encoding* agar dapat diproses oleh algoritma *clustering* berbasis jarak. Data hasil transformasi kemudian dinormalisasi menggunakan *RobustScaler*, yaitu metode normalisasi berbasis median dan *interquartile range* yang lebih tahan terhadap *outlier* dibandingkan *Min-Max Scaling* atau *StandardScaler*. Setelah normalisasi, dilakukan reduksi dimensi menggunakan *Principal Component Analysis* (PCA) dengan parameter *n_components=2* untuk menyederhanakan struktur data sekaligus mempermudah proses *clustering* dan visualisasi hasil.

2.4 Penetapan Metode Clustering

Penentuan jumlah *cluster* dilakukan dengan menjalankan setiap algoritma pada variasi $k = 2$ hingga $k = 7$, kemudian membandingkan nilai *Silhouette Score* dan *Davies-Bouldin Index* (DBI) pada setiap variasi tersebut. Meskipun $k = 2$ pada beberapa metode menghasilkan nilai evaluasi tertinggi, penelitian ini menetapkan jumlah *cluster* sebanyak tiga ($k = 3$) agar sesuai dengan tujuan penelitian, yaitu mengelompokkan data ke dalam tiga kategori rekomendasi penerima bantuan pendidikan: sangat layak, layak, dan tidak layak menerima bantuan.

Lima metode *clustering* diterapkan pada dataset yang identik agar perbedaan hasil semata-mata mencerminkan karakteristik algoritma, bukan perbedaan data. *K-Means* diterapkan dengan parameter *n_clusters* = *k* dan perhitungan jarak menggunakan *Euclidean Distance* terhadap *centroid*. *Fuzzy C-Means* (FCM) diterapkan dengan parameter tingkat *fuzziness m* = 2,0, iterasi maksimum *maxiter* = 1000, dan toleransi *error* = 0,005, dengan penentuan *cluster* dominan melalui fungsi *argmax* pada matriks keanggotaan. *Gaussian Mixture Model* (GMM) menggunakan parameter *n_components* = *k* dengan jenis kovarians *covariance_type* = 'full'. *Spectral Clustering* diterapkan dengan parameter *n_clusters* = *k*, *affinity* = 'nearest_neighbors' untuk membangun matriks kemiripan antartetangga terdekat, serta *random_state* = 42 agar hasil dapat direproduksi; proses dilanjutkan dengan dekomposisi eigen terhadap matriks Laplacian sebelum pengelompokan. *Agglomerative Hierarchical Clustering* diterapkan dengan parameter *n_clusters* = *k*, metode *linkage* = 'ward', dan metrik jarak Euclidean, dengan proses penggabungan cluster dilakukan secara bertahap berdasarkan tingkat kemiripan hingga terbentuk tiga *cluster*.

2.5 Evaluasi Perbandingan

Kualitas hasil pengelompokan dari kelima metode dievaluasi menggunakan dua metrik evaluasi internal, yaitu *Silhouette Score*, yang mengukur tingkat kohesi dan separasi *cluster* dengan rentang nilai −1 hingga 1, dan *Davies-Bouldin Index*, yang mengukur rasio jarak intra-*cluster* terhadap jarak inter-*cluster*, semakin rendah nilainya

DOI: https://doi.org/10.31004/riggs.v5i2.11894

semakin baik kualitas *cluster* yang dihasilkan. Metode dengan *Silhouette Score* tertinggi dan DBI terendah pada *k* = 3 ditetapkan sebagai metode dengan performa terbaik untuk mendukung sistem pengambilan keputusan penentuan penerima bantuan pendidikan. Seluruh proses analisis dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka Pandas, NumPy, Scikit-Learn, dan Scikit-Fuzzy.

Seluruh eksperimen dijalankan pada lingkungan Google Colaboratory tanpa akselerasi GPU, mengingat kelima algoritma clustering yang digunakan bersifat berbasis CPU dan tidak memerlukan pemrosesan paralel berskala besar. Setiap metode dijalankan pada kondisi praproses data yang identik, dengan parameter acak (random_state) yang tetap pada metode yang membutuhkannya, agar hasil evaluasi dapat direproduksi pada percobaan berikutnya. Konsistensi kondisi pengujian ini penting agar perbedaan performa yang teramati benar-benar mencerminkan karakteristik masing-masing algoritma, bukan akibat variasi acak pada tahap inisialisasi maupun perbedaan perlakuan data antarmetode.

2.6 Kesimpulan dan Penentuan Metode Terbaik

Tahap akhir penelitian meliputi penarikan kesimpulan dan penyusunan rekomendasi berdasarkan seluruh proses penelitian, mulai dari identifikasi masalah, pra-pemrosesan data, implementasi metode *clustering*, hingga evaluasi hasil. Kesimpulan disusun berdasarkan nilai *Silhouette Score* dan (DBI) untuk menentukan metode *clustering* dengan performa terbaik secara objektif. Selain itu, penelitian mengidentifikasi karakteristik cluster yang terbentuk sebagai dasar pengelompokan siswa berdasarkan kondisi sosial ekonomi dan akses pendidikan. Hasil penelitian menunjukkan bahwa penerapan metode clustering dapat mendukung proses seleksi bantuan pendidikan yang lebih objektif, transparan, dan berbasis data.

3. Hasil dan Diskusi

3.1 Hasil

Pemeriksaan awal terhadap dataset siswa SMKN 1 Lambu (1.241 data, 15 variabel) dan dataset Academic Performance Retention (574 data, 17 variabel) menunjukkan bahwa kedua dataset tidak mengandung nilai kosong maupun data duplikat, sehingga dapat langsung digunakan pada tahap pra-pemrosesan selanjutnya. Atribut kategorikal dikonversi menjadi bentuk numerik melalui teknik *encoding*, kemudian diseragamkan skalanya menggunakan *RobustScaler*, dan direduksi menjadi dua komponen utama (PCA 1 dan PCA 2) menggunakan *Principal Component Analysis* untuk mempermudah proses *clustering* dan visualisasi.

Pengujian jumlah *cluster* pada rentang *k* = 2 hingga *k* = 7 menunjukkan bahwa nilai evaluasi terbaik pada sebagian besar metode diperoleh pada *k* = 2. Namun demikian, penelitian ini menetapkan *k* = 3 sesuai tujuan penelitian, yaitu menghasilkan tiga kategori rekomendasi penerima bantuan pendidikan: sangat layak, layak, dan tidak layak menerima bantuan. Hasil evaluasi kelima metode pada *k* = 3 untuk dataset utama (siswa SMKN 1 Lambu) dirangkum pada Tabel 1.

Tabel 1. Hasil Evaluasi Clustering pada k = 3 (Dataset Utama)

Metode	Silhouette Score	Davies-Bouldin Index
Agglomerative Hierarchical Clustering	0,652852	0,597686
K-Means	0,623140	0,623611
Fuzzy C-Means	0,572057	0,692978
Gaussian Mixture Model	0,310727	1,131143
Spectral Clustering	0,092729	1,145970

Pada dataset pembanding (Academic Performance Retention), hampir seluruh metode kecuali Spectral Clustering menghasilkan Silhouette Score di atas 0,98 dan Davies-Bouldin Index di bawah 0,36 pada *k* = 3, dengan Fuzzy C-Means mencatatkan nilai tertinggi sebesar 0,995794. Spectral Clustering tetap konsisten menghasilkan performa terendah pada kedua dataset, bahkan bernilai negatif pada dataset pembanding (-0,802913 pada *k* = 2).

Analisis nilai rata-rata atribut pada setiap cluster menunjukkan bahwa jarak tempuh ke sekolah menjadi atribut pembeda paling dominan pada seluruh metode. Pada hasil *Agglomerative Hierarchical Clustering*, terbentuk satu cluster dengan rata-rata jarak tempuh 19,38 km disertai jumlah saudara yang relatif banyak, satu cluster dengan jarak tempuh dan tanggungan keluarga jauh lebih rendah, serta satu cluster dengan karakteristik menengah (rata-rata 9,65 km). Pola pembeda serupa juga tampak pada K-Means dan Fuzzy C-Means, sedangkan pada Spectral

DOI: https://doi.org/10.31004/riggs.v5i2.11894

Clustering, visualisasi hasil reduksi PCA memperlihatkan titik data yang sebagian besar bertumpuk pada satu wilayah dan tidak membentuk batas cluster yang jelas.

Perbedaan pola sebaran antarmetode tersebut turut terlihat pada bentuk visual hasil reduksi PCA. Pada Agglomerative Hierarchical Clustering, K-Means, dan Fuzzy C-Means, ketiga cluster tampak terpisah dengan batas yang relatif jelas meskipun terdapat sedikit tumpang tindih di area perbatasan, sementara pada Gaussian Mixture Model batas antarcluster tampak lebih menyebar mengikuti bentuk elips sesuai asumsi distribusinya. Perbandingan pada dataset Academic Performance Retention menunjukkan pola yang relatif konsisten, di mana metode yang unggul pada dataset utama juga cenderung memberikan nilai evaluasi yang baik pada dataset kedua, kecuali Spectral Clustering yang tetap menunjukkan performa terendah pada kedua dataset. Konsistensi ini mengindikasikan bahwa karakteristik algoritma, bukan semata-mata karakteristik dataset, menjadi faktor utama yang menentukan kualitas hasil pengelompokan pada penelitian ini.

3.2 Diskusi

Urutan performa pada Tabel 1 mencerminkan bagaimana setiap algoritma menangani karakteristik data sosial ekonomi yang digunakan. *Agglomerative Hierarchical Clustering* unggul karena bekerja dengan menggabungkan data secara bertahap berdasarkan tingkat kemiripan (*linkage* = ward), sehingga struktur hubungan antardata terbentuk secara alami tanpa bergantung pada penentuan titik pusat awal yang bersifat acak. Pendekatan bertahap ini membuatnya lebih tahan terhadap distribusi data yang tidak simetris, sebagaimana terlihat pada atribut jarak tempuh dan pendapatan orang tua yang cenderung menyebar tidak merata. K-Means menempati posisi kedua karena, meskipun bergantung pada centroid dan sensitif terhadap inisialisasi awal, kesederhanaan mekanisme jarak Euclidean-nya tetap efektif ketika sebagian besar atribut telah dinormalisasi. Hasil ini juga sejalan dengan penelitian yang membandingkan K-Means, K-Medoids, dan Fuzzy C-Means, di mana K-Means tetap menunjukkan performa yang kompetitif pada data yang telah melalui proses pra-pemrosesan dengan baik [13]. Fuzzy C-Means sedikit di bawah K-Means karena konsep derajat keanggotaannya justru membuat batas antarcluster menjadi lebih kabur pada data yang sebenarnya memiliki pemisahan cukup tegas, sehingga sifat *soft clustering*-nya tidak sepenuhnya menjadi keunggulan pada kasus ini. Temuan tersebut berbeda dengan penelitian pada data status sosial ekonomi keluarga penerima manfaat yang menunjukkan bahwa Fuzzy C-Means mampu memberikan hasil yang lebih baik ketika batas antar kelompok relatif tidak tegas[14] .

Performa Gaussian Mixture Model yang jauh lebih rendah berkaitan dengan asumsi dasarnya bahwa data mengikuti distribusi Gaussian. Atribut sosial ekonomi pada penelitian ini, khususnya jarak tempuh dan jumlah saudara, tidak sepenuhnya berdistribusi normal, sehingga model probabilistik ini kesulitan menempatkan batas keanggotaan cluster secara tegas. Hasil serupa juga ditemukan pada penelitian yang membandingkan K-Means dan Gaussian Mixture Model, di mana performa GMM sangat dipengaruhi oleh karakteristik distribusi data yang digunakan [15]. Kondisi paling ekstrem terjadi pada Spectral Clustering: nilai Silhouette Score yang mendekati nol dan DBI tertinggi mengindikasikan bahwa pendekatan berbasis graf dan matriks kemiripan *nearest_neighbors* kurang sesuai untuk data tabular sosial ekonomi yang bersifat diskret dan tidak memiliki struktur relasi antarobjek yang kompleks, berbeda dengan jenis data yang biasanya menjadi keunggulan metode ini. Sebaliknya, penelitian lain menunjukkan bahwa Spectral Clustering mampu menghasilkan performa yang baik pada data dengan struktur hubungan yang kompleks, sehingga efektivitasnya sangat dipengaruhi oleh karakteristik dataset [16].

Tingginya nilai evaluasi pada dataset pembanding tidak menunjukkan superioritas metode tertentu, melainkan mengindikasikan bahwa dataset tersebut memiliki struktur yang jauh lebih homogen sehingga hampir semua algoritma dapat menemukan pemisahan tersebut dengan mudah. Perbandingan dua dataset ini memberi wawasan penting: kualitas clustering yang sangat tinggi pada data yang homogen tidak lantas menjadikannya tolok ukur metode terbaik, karena tujuan penelitian tetap berpusat pada dataset utama yang mencerminkan kondisi lapangan sesungguhnya, yaitu data siswa dengan variasi sosial ekonomi yang kompleks.

Pola pembeda berbasis jarak tempuh dan jumlah saudara konsisten dengan premis penelitian bahwa akses fisik terhadap sekolah dan beban tanggungan keluarga berkorelasi dengan tingkat kebutuhan bantuan pendidikan. Sementara itu, kemampuan Spectral Clustering membedakan karakteristik cluster secara deskriptif ternyata tidak sejalan dengan kualitas pemisahan yang terukur secara statistik, menegaskan bahwa interpretasi deskriptif saja tidak cukup untuk menilai kualitas sebuah metode *clustering*.

Temuan bahwa Agglomerative Hierarchical Clustering memberikan hasil terbaik memperkuat sekaligus memperluas temuan [12], yang melaporkan metode ini dengan pendekatan *average linkage* menghasilkan Silhouette Score sebesar 0,79 pada kasus pengelompokan prioritas penerima bantuan rumah. Meskipun nilai pada penelitian ini (0,652852) sedikit lebih rendah, kedua penelitian sama-sama menunjukkan keunggulan pendekatan hierarkis pada data sosial ekonomi dengan variasi tinggi antarindividu, berbeda dengan penelitian yang membandingkan Fuzzy C-Means dan Gaussian Mixture Model pada data kesejahteraan masyarakat Indonesia,

DOI: https://doi.org/10.31004/riggs.v5i2.11894

yang menyimpulkan FCM lebih unggul pada data dengan batas kategori tidak tegas [8]. Perbedaan ini mengindikasikan bahwa keunggulan suatu metode *clustering* sangat bergantung pada karakteristik atribut dan tingkat tumpang tindih data, bukan bersifat mutlak berlaku pada seluruh konteks data pendidikan atau sosial ekonomi.

Penelitian ini juga memperluas temuan [7], yang membandingkan K-Means dan Fuzzy C-Means untuk pengelompokan indikator pendidikan tingkat kabupaten/kota, dengan menambahkan tiga metode pembanding sekaligus dua metrik evaluasi yang saling melengkapi. Penggunaan Silhouette Score dan Davies-Bouldin Index secara bersamaan juga sejalan dengan penelitian benchmarking indeks validitas clustering yang merekomendasikan penggunaan lebih dari satu metrik evaluasi agar kualitas cluster dapat dinilai secara lebih komprehensif [17]. Hasil penelitian ini menunjukkan bahwa K-Means, meski sederhana, tetap kompetitif dan hanya sedikit berada di bawah metode hierarkis, sehingga tetap layak menjadi alternatif ketika efisiensi komputasi menjadi pertimbangan utama.

Sementara itu, rendahnya performa Spectral Clustering pada penelitian ini berbeda dengan temuan [11], yang melaporkan bahwa Spectral Clustering meningkatkan kualitas pengelompokan pengguna pada sistem rekomendasi. Perbedaan ini wajar mengingat sistem rekomendasi umumnya bekerja pada data relasional atau data dengan struktur graf yang eksplisit, sedangkan data sosial ekonomi pada penelitian ini bersifat atributif tanpa relasi antarobjek yang dapat dimanfaatkan secara optimal oleh pendekatan berbasis matriks kemiripan. Temuan ini memperjelas batas penerapan Spectral Clustering: berpotensi unggul pada data berstruktur relasi kompleks, tetapi kurang sesuai untuk data tabular sosial ekonomi seperti pada penelitian ini.

Implikasinya, pemilihan metode clustering dalam pengembangan sistem pendukung keputusan tidak dapat dilakukan secara generik, melainkan perlu mempertimbangkan tipe data yang tersedia di lapangan. Data administratif sekolah pada umumnya bersifat tabular dan atributif tanpa relasi eksplisit antarsiswa, sehingga metode berbasis centroid atau hierarki lebih sesuai diadopsi dibandingkan metode berbasis graf seperti Spectral Clustering, yang keunggulannya baru terlihat pada data dengan struktur relasi yang eksplisit dan kompleks.

Dibandingkan dengan pendekatan Sistem Pendukung Keputusan berbasis *Profile Matching* yang digunakan [5] untuk menentukan penerima beasiswa, pendekatan berbasis *clustering* pada penelitian ini menawarkan keunggulan berupa proses pengelompokan yang tidak memerlukan penetapan bobot atau kriteria kelayakan secara manual di awal, sehingga potensi bias akibat penentuan bobot subjektif dapat diminimalkan. Secara keseluruhan, hasil penelitian ini memperkuat posisi Agglomerative Hierarchical Clustering sebagai metode yang direkomendasikan untuk mendukung Sistem Pendukung Keputusan penentuan penerima bantuan pendidikan, dengan catatan bahwa keunggulan ini bersifat spesifik terhadap karakteristik data sosial ekonomi yang memiliki variasi tinggi dan tidak sepenuhnya berdistribusi normal.

Dari sisi penerapan, hasil pengelompokan berbasis Agglomerative Hierarchical Clustering dapat diintegrasikan ke dalam sistem administrasi sekolah dalam bentuk modul rekomendasi yang berjalan setiap awal tahun ajaran, ketika data sosial ekonomi siswa baru maupun pembaruan data siswa lama tersedia. Label cluster yang dihasilkan tetap perlu diverifikasi oleh pihak sekolah sebelum keputusan akhir diambil, mengingat clustering bersifat tidak terawasi dan tidak memiliki mekanisme validasi terhadap kondisi lapangan yang sesungguhnya. Dengan demikian, peran algoritma lebih tepat diposisikan sebagai alat bantu pengambilan keputusan yang mempersempit dan mengobjektifkan proses seleksi, bukan sebagai pengganti sepenuhnya penilaian pihak sekolah.

4. Kesimpulan

Penelitian ini berhasil menerapkan dan membandingkan kinerja lima metode *clustering*, yaitu K-Means, Fuzzy C-Means, Gaussian Mixture Model, Spectral Clustering, dan Agglomerative Hierarchical Clustering, dalam mengelompokkan calon penerima bantuan pendidikan berdasarkan kondisi sosial ekonomi siswa. Pada dataset utama, yaitu data siswa SMKN 1 Lambu, Agglomerative Hierarchical Clustering terbukti memberikan kualitas pengelompokan terbaik pada jumlah cluster tiga ($k = 3$) dengan Silhouette Score sebesar 0,652852 dan Davies-Bouldin Index sebesar 0,597686, mengungguli K-Means, Fuzzy C-Means, Gaussian Mixture Model, dan Spectral Clustering. Pengujian pada dataset pembanding, Academic Performance Retention, menunjukkan bahwa hampir seluruh metode mampu menghasilkan kualitas cluster yang sangat baik pada data yang lebih homogen, kecuali Spectral Clustering yang secara konsisten menghasilkan performa terendah pada kedua dataset. Temuan ini menjawab rumusan masalah penelitian dengan menunjukkan bahwa performa metode *clustering* sangat dipengaruhi oleh karakteristik data yang digunakan, dan bahwa pendekatan hierarkis lebih andal dalam menangani data sosial ekonomi yang memiliki variasi tinggi serta distribusi yang tidak sepenuhnya normal dibandingkan pendekatan berbasis centroid, probabilistik, maupun graf. Secara praktis, hasil penelitian ini menunjukkan bahwa jarak tempuh ke sekolah dan jumlah tanggungan keluarga merupakan atribut yang paling berpengaruh dalam membedakan tingkat kebutuhan bantuan pendidikan, sehingga dapat dijadikan pertimbangan utama oleh pihak sekolah dalam merancang kriteria seleksi. Rekomendasi penggunaan Agglomerative Hierarchical Clustering

DOI: https://doi.org/10.31004/riggs.v5i2.11894

sebagai dasar Sistem Pendukung Keputusan diharapkan dapat menggantikan proses seleksi manual yang selama ini rentan terhadap subjektivitas, sehingga distribusi bantuan pendidikan menjadi lebih objektif, konsisten, dan tepat sasaran. Penelitian selanjutnya disarankan memperluas cakupan data pada beberapa sekolah atau lembaga pendidikan sekaligus menambahkan indikator lain, seperti kepemilikan aset keluarga dan prestasi akademik, agar hasil pengelompokan lebih komprehensif. Perbandingan dengan metode *clustering* lain, seperti DBSCAN atau Self-Organizing Map, serta penambahan metrik evaluasi seperti Calinski-Harabasz Index, juga dapat dipertimbangkan untuk memperkaya hasil evaluasi pada penelitian mendatang. Kontribusi utama penelitian ini terletak pada penyediaan bukti komparatif langsung antara lima metode clustering pada data seleksi bantuan pendidikan yang sebelumnya belum banyak dikaji secara bersamaan dalam satu kerangka evaluasi. Temuan ini diharapkan dapat menjadi rujukan bagi sekolah maupun instansi penyelenggara program bantuan pendidikan lain yang memiliki karakteristik data serupa, terutama dalam menimbang antara kesederhanaan implementasi, kebutuhan interpretasi hasil, dan tingkat akurasi pengelompokan yang diinginkan. Keterbatasan penelitian ini, yaitu cakupan data yang masih berasal dari satu sekolah dan periode pengumpulan yang terbatas, perlu menjadi perhatian dalam menggeneralisasi hasil pada konteks yang lebih luas. Dari perspektif pengembangan sistem informasi sekolah, hasil penelitian ini membuka peluang integrasi modul clustering ke dalam aplikasi manajemen kesiswaan yang sudah berjalan, sehingga proses pembaruan rekomendasi penerima bantuan dapat dilakukan secara berkala tanpa memerlukan pengolahan data manual dari awal setiap periode. Penerapan seperti ini sejalan dengan arah digitalisasi layanan pendidikan yang semakin banyak diadopsi oleh sekolah maupun dinas pendidikan di berbagai daerah, di mana ketersediaan data siswa yang terstruktur menjadi modal penting untuk mendukung pengambilan keputusan berbasis bukti. Meskipun demikian, keberhasilan penerapan sistem semacam ini tidak hanya bergantung pada ketepatan algoritma yang digunakan, tetapi juga pada kualitas data masukan, konsistensi pembaruan data, serta kesiapan sumber daya manusia di sekolah dalam mengoperasikan dan menginterpretasikan hasil sistem. Oleh karena itu, pengembangan lebih lanjut disarankan turut mempertimbangkan aspek pelatihan bagi operator sekolah dan penyusunan panduan teknis penggunaan sistem, agar manfaat pendekatan berbasis data mining ini dapat dirasakan secara berkelanjutan dan tidak bergantung pada satu pihak saja. Penelitian lanjutan juga dapat diarahkan pada pengujian stabilitas hasil clustering terhadap perubahan komposisi data dari waktu ke waktu, mengingat kondisi sosial ekonomi siswa berpotensi berubah antarperiode akibat faktor eksternal seperti perubahan pekerjaan orang tua atau relokasi tempat tinggal. Evaluasi semacam ini penting untuk memastikan bahwa rekomendasi yang dihasilkan sistem tetap relevan dan tidak bias terhadap pola data pada satu periode pengumpulan saja. Selain itu, penggabungan pendekatan clustering dengan metode klasifikasi terawasi pada tahap lanjutan berpotensi menghasilkan sistem yang lebih adaptif, misalnya dengan memanfaatkan label cluster sebagai dasar pelatihan model prediktif untuk periode berikutnya. Pendekatan hybrid semacam ini belum banyak dieksplorasi pada konteks seleksi bantuan pendidikan berbasis kondisi sosial ekonomi, sehingga berpotensi menjadi arah penelitian yang relevan untuk dikembangkan lebih lanjut, baik dari sisi akurasi maupun efisiensi proses evaluasi yang dilakukan pihak sekolah. Perlu dicatat pula bahwa evaluasi pada penelitian ini sepenuhnya bersandar pada metrik internal, yaitu Silhouette Score dan Davies-Bouldin Index, yang mengukur kualitas pemisahan cluster tanpa mempertimbangkan label kebenaran lapangan mengenai kelayakan penerima bantuan. Validasi eksternal, misalnya melalui perbandingan hasil cluster dengan keputusan seleksi yang selama ini dilakukan pihak sekolah atau dengan indikator kesejahteraan keluarga dari sumber lain, akan memperkuat keyakinan terhadap kualitas rekomendasi yang dihasilkan sistem. Ketiadaan validasi eksternal pada penelitian ini merupakan keterbatasan yang perlu ditindaklanjuti pada penelitian selanjutnya, agar hasil pengelompokan tidak hanya baik secara statistik, tetapi juga terbukti relevan secara substansi di lapangan. Secara keseluruhan, penelitian ini menegaskan bahwa pemilihan metode clustering bukan sekadar persoalan teknis, melainkan keputusan yang perlu disesuaikan dengan karakteristik data dan konteks penggunaannya. Pada kasus seleksi bantuan pendidikan berbasis kondisi sosial ekonomi, pendekatan hierarkis terbukti memberikan hasil yang paling andal dibandingkan pendekatan berbasis centroid, probabilistik, maupun graf, sehingga layak dipertimbangkan sebagai fondasi awal pengembangan sistem pendukung keputusan di lingkungan sekolah. Ke depan, kolaborasi antara pihak sekolah, pengembang sistem, dan peneliti diharapkan dapat mempercepat transisi dari proses seleksi manual menuju proses yang lebih terukur, transparan, dan berkelanjutan, tanpa menghilangkan peran penting pertimbangan manusia dalam pengambilan keputusan akhir. Dengan mempertimbangkan seluruh temuan dan keterbatasan yang telah diuraikan, penelitian ini diharapkan dapat memberikan kontribusi nyata, baik secara akademis melalui perbandingan komparatif lima metode clustering, maupun secara praktis melalui rekomendasi metode yang dapat diadopsi pihak sekolah dalam merancang sistem seleksi bantuan pendidikan yang lebih objektif dan akuntabel.

Referensi

[1] I. A. Muzaki, Mujahidah, and M. Erihadiana, “Manajemen sumber daya manusia sebagai basis penguatan kualitas pendidikan,” *Muntazam: J. Manaj. Pendidik. Islam*, vol. 2, no. 2, pp. 14–31, 2021, doi: 10.35706/muntazam.v2i2.5998.

DOI: https://doi.org/10.31004/riggs.v5i2.11894

[2] A. Abrianto, "Application of K-Means clustering algorithm for determining PIP scholarship recipients at SMPN 9 Blitar," *JOSAR (J. Students Acad. Res.)*, vol. 9, no. 1, pp. 204–214, 2024, doi: 10.35457/josar.v9i1.3105.
[3] D. Agustini, M. Farida, M. Sari, and M. E. Rosadi, "Rancang bangun sistem informasi beasiswa (studi kasus: Uniska MAB Banjarmasin)," *Technologia: J. Ilm.*, vol. 13, no. 3, p. 270, 2022, doi: 10.31602/tji.v13i3.7558.
[4] A. R. Nisa *et al.*, "Prediction on target of underprivileged scholarships using fuzzy logic method," *J. Appl. Sci. Technol. Humanit.*, vol. 1, no. 2, pp. 159–173, 2024, doi: 10.62535/4bg52465.
[5] A. Setiyowati, L. A. Ramadhani, and M. K. Amin, "Sistem pendukung keputusan menentukan penerima beasiswa kurang mampu menggunakan metode Profile Matching," *J. Inform. Upgris*, vol. 6, no. 1, 2020, doi: 10.26877/jiu.v6i1.4896.
[6] K. A. Mustari, P. Assiroj, B. Hartati, and F. Samuel, "Implementasi data mining pada instansi pemerintahan," *JATI (J. Mhs. Tek. Inform.)*, vol. 8, no. 3, pp. 3137–3142, 2024, doi: 10.36040/jati.v8i3.9618.
[7] 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.
[8] E. W. Ambarsari, N. Dwitiyanti, N. Selvia, W. N. Cho, and P. D. Mardika, "Comparison approaches of the Fuzzy C-Means and Gaussian Mixture Model in clustering the welfare of the Indonesian people," *KnE Social Sciences*, vol. 8, no. 9, pp. 16–22, 2023, doi: 10.18502/kss.v8i9.13315.
[9] A. S. Muthahharah, M. A. Tiro, and A. Aswi, "Application of soft-clustering analysis using expectation maximization algorithms on Gaussian Mixture Model," *J. Varian*, vol. 6, no. 1, pp. 71–80, 2022, doi: 10.30812/varian.v6i1.2142.
[10] M. R. Gusmansyah, Rahmaddeni, Rohid, S. Daulay, and A. Rivaldi, "Perbandingan metode machine learning untuk klastering penerima bantuan pendidikan siswa," *J. Pustaka AI*, vol. 5, no. 2, pp. 193–203, 2025, doi: 10.55382/jurnalpustakaai.v5i2.1139.
[11] M. Muhammad, "Recommendation system using user-based collaborative filtering and spectral clustering," in *Proceedings of EAI*, 2021. doi: 10.4108/eai.17-7-2021.2312409.
[12] K. Y. Katrina and J. B. B. Darmawan, "Penerapan algoritma Agglomerative Hierarchical untuk clustering peringkat prioritas penerima bantuan rumah," in *Semin. Nas. Tek. Elektro, Sist. Inf., dan Tek. Inform. (SNESTIK V)*, 2025, pp. 105–111.
[13] H. Syukron, M. F. Fayyad, and F. J. Fauzan, "Comparison K-Means, K-Medoids and Fuzzy C-Means for clustering customer data with LRFM model," *Malcom: Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, pp. 76–83, 2022.
[14] P. S. Saputra, G. R. Dantes, and I. G. A. Gunadi, "Perbandingan algoritma Fuzzy C-Means dan algoritma Naive Bayes dalam menentukan Keluarga Penerima Manfaat (KPM) berdasarkan status sosial ekonomi (SSE) terendah," *JST (J. Sains dan Teknol.)*, vol. 10, no. 1, pp. 64–74, 2021, doi: 10.23887/jstundiksha.v10i1.23340.
[15] R. Rahmattullah, Indwiarti, and A. A. Rohmawati, "Clustering harga rumah: perbandingan model K-Means dan Gaussian Mixture Model," *e-Proceeding of Eng.*, vol. 10, no. 3, pp. 3441–3449, 2023.
[16] F. A. Totti and N. Setiyawati, "A comparative study of K-Means, Gaussian Mixture Model, and Spectral Clustering for facial emotion recognition," *Sistemasi*, vol. 14, no. 6, p. 3007, 2025, doi: 10.32520/stmsi.v14i6.5668.
[17] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Benchmarking validity indices for evolutionary K-means clustering performance," *Sci. Rep.*, vol. 15, p. 21842, 2025, doi: 10.1038/s41598-025-08473-6.

DOI: https://doi.org/10.31004/riggs.v5i2.11894