



K-Means Clustering of Higher Education Resources Across Indonesian Provinces in 2025

Klasterisasi Sumber Daya Pendidikan Tinggi Antarprovinsi di Indonesia Tahun 2025 Menggunakan K-Means

Kristia Yulilawan^{1*}, Agung Hardiansyah²

^{1,2}Program Studi Informatika, STMIK Pesat Nabire, Indonesia

E-Mail: ¹christianpesat@gmail.com, ²syagungelhaq12mail@gmail.com

Received May 18th 2026; Revised Jun 23 th 2026; Accepted Jul 05th 2026; Available Online Jun 28th 2026

Corresponding Author: Kristia Yulilawan

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The unequal distribution of higher education institutions and academic resources across Indonesian provinces affects the equity and quality of higher education services. Using 2025 data, this study clusters 38 Indonesian provinces using the K-Means algorithm implemented in Orange Data Mining based on the number of higher education institutions, educators, students, student-to-educator ratio, and educator requirements. The data were preprocessed, standardized using z-score normalization, grouped into three clusters, and evaluated using the Silhouette Score. The resulting average silhouette value was 0.520, indicating a reasonably well-separated cluster structure. Cluster 1 contained provinces with relatively limited higher education capacity, Cluster 2 represented provinces with educator-capacity pressure, and Cluster 3 comprised major national academic centers. The novelty of this study lies in integrating institutional scale, academic human resources, student volume, and staffing adequacy within a single province-level clustering model. The findings provide an empirical basis for differentiated policies on institution expansion, educator allocation, and regional higher education capacity development.

Keywords: Data Mining, Higher Education, K-Means Clustering, Orange Data Mining, Resource Distribution

Abstrak

Ketimpangan distribusi perguruan tinggi dan sumber daya pendidikan tinggi di Indonesia memengaruhi pemerataan akses dan kualitas layanan pendidikan tinggi. Dengan menggunakan data tahun 2025, penelitian ini mengelompokkan 38 provinsi menggunakan algoritma K-Means pada Orange Data Mining berdasarkan jumlah perguruan tinggi, tenaga pendidik, mahasiswa, rasio mahasiswa terhadap tenaga pendidik, dan kebutuhan tenaga pendidik. Data melalui tahap pra-pemrosesan, standarisasi z-score, pembentukan tiga cluster, dan evaluasi menggunakan Silhouette Score. Nilai silhouette rata-rata sebesar 0,520 menunjukkan struktur cluster yang cukup baik dan terpisah. Cluster 1 menggambarkan provinsi dengan kapasitas pendidikan tinggi yang relatif terbatas, Cluster 2 menunjukkan tekanan kebutuhan tenaga pendidik, dan Cluster 3 mencakup pusat-pusat aktivitas akademik nasional. Kebaruan penelitian terletak pada integrasi skala institusi, sumber daya akademik, jumlah mahasiswa, dan kecukupan tenaga pendidik dalam satu model clustering tingkat provinsi. Hasil penelitian ini menyediakan dasar empiris bagi kebijakan diferensial terkait pengembangan institusi, distribusi tenaga pendidik, dan penguatan kapasitas pendidikan tinggi.

Kata Kunci: Data Mining, Distribusi Sumber Daya, K-Means Clustering, Orange Data Mining, Pendidikan Tinggi

1. PENDAHULUAN

Pendidikan tinggi memiliki peran penting dalam meningkatkan kualitas sumber daya manusia, daya saing wilayah, serta pembangunan nasional. Dataset penelitian memperlihatkan ketimpangan yang lebar: jumlah perguruan tinggi antarprovinsi berkisar dari 7 hingga 348 institusi, jumlah tenaga pendidik dari 98 hingga 24.179 orang, dan jumlah mahasiswa dari 3.347 hingga 671.153 orang. Rasio mahasiswa terhadap tenaga pendidik juga bervariasi antara 19 dan 40 mahasiswa per pendidik. Rentang tersebut menunjukkan bahwa kapasitas layanan pendidikan tinggi masih terkonsentrasi pada provinsi tertentu, sementara sebagian wilayah menghadapi keterbatasan institusi dan sumber daya akademik. Oleh karena itu, diperlukan analisis

berbasis data yang tidak hanya mempertimbangkan jumlah perguruan tinggi, tetapi juga tenaga pendidik, mahasiswa, rasio layanan akademik, serta kebutuhan tenaga pendidik.

Pemanfaatan data pendidikan semakin penting untuk mendukung pengambilan keputusan yang terukur. Yunita et al. [1] menegaskan bahwa ekosistem data pendidikan tinggi di Indonesia berpotensi besar untuk mendukung integrasi informasi dan kebijakan berbasis data. Dalam analisis pendidikan, metode data mining, khususnya clustering, dapat digunakan untuk mengelompokkan wilayah berdasarkan kemiripan karakteristik. Putri et al. [2] menggunakan K-Means untuk mengelompokkan provinsi di Indonesia berdasarkan indikator pendidikan, sedangkan Mustakim dan Kamal [3] menerapkan K-Means untuk mengklasifikasikan manajemen mutu pendidikan menengah. Widianoro et al. [4] juga menggunakan K-Means untuk mengelompokkan provinsi berdasarkan indikator pendidikan SMA/ sederajat. Penelitian tersebut menunjukkan bahwa K-Means efektif digunakan untuk menganalisis variasi karakteristik pendidikan berdasarkan wilayah.

Kajian terdahulu menunjukkan bahwa clustering relevan untuk analisis pendidikan, tetapi memiliki fokus dan keterbatasan yang berbeda. Arfan dan Pekei [5] menggunakan K-Means untuk mengevaluasi distribusi guru, siswa, dan sekolah; kelebihanannya terletak pada interpretasi kewilayahan, tetapi unit analisisnya masih pada pendidikan sekolah. Talib et al. [6] menggabungkan K-Means dan Support Vector Machine untuk mengidentifikasi perilaku mahasiswa, sehingga mampu menghubungkan segmentasi dan klasifikasi, namun tidak membahas ketimpangan kapasitas antardaerah. Rani et al. [7] mengombinasikan K-Means dan FP-Growth untuk menemukan pola subjek mahasiswa, sedangkan Mohamed Nafuri et al. [8] membandingkan K-Means, BIRCH, dan DBSCAN dalam performa akademik; kedua penelitian tersebut menekankan karakteristik individu, bukan distribusi institusi dan sumber daya regional. Ikotun et al. [9] menegaskan efisiensi dan kemudahan interpretasi K-Means, tetapi juga mengingatkan akan sensitivitasnya terhadap skala data, centroid awal, dan pencilan. Corradini et al. [10] menunjukkan keunggulan Orange Data Mining untuk membangun alur analisis visual yang transparan, meskipun kualitas hasil tetap bergantung pada pra-pemrosesan dan pemilihan parameter. Analisis kritis tersebut menunjukkan perlunya normalisasi, evaluasi kualitas cluster, dan interpretasi berbasis konteks dalam penelitian ini.

Meskipun K-Means telah banyak digunakan dalam bidang pendidikan, sebagian besar penelitian sebelumnya berfokus pada pendidikan dasar dan menengah, perilaku mahasiswa, kinerja akademik, atau distribusi guru-siswa. Penelitian tersebut umumnya menggunakan unit analisis individu atau sekolah dan belum mengintegrasikan skala institusi, jumlah tenaga pendidik, jumlah mahasiswa, rasio layanan, dan kebutuhan tenaga pendidik di seluruh provinsi di Indonesia. Berbeda dari penelitian sebelumnya, studi ini menggunakan 38 provinsi sebagai unit analisis dan membangun tipologi kapasitas pendidikan tinggi melalui kombinasi indikator institusional dan sumber daya akademik. Perbedaan tersebut memperluas pemanfaatan educational data mining dari prediksi atau segmentasi individu menuju analisis kebijakan pendidikan tinggi berbasis wilayah.

Berdasarkan celah tersebut, penelitian ini menerapkan algoritma K-Means berbasis Orange Data Mining untuk mengelompokkan 38 provinsi berdasarkan karakteristik distribusi perguruan tinggi dan sumber daya pendidikan tinggi. Penelitian bertujuan mengidentifikasi pola kemiripan antarprovinsi, menjelaskan karakteristik setiap cluster, dan menilai kualitas pemisahan cluster menggunakan Silhouette Score. Kebaruan penelitian terletak pada integrasi lima indikator kapasitas pendidikan tinggi dalam satu model clustering tingkat provinsi. Secara praktis, hasil penelitian dapat digunakan untuk menyusun kebijakan yang berbeda menurut karakteristik cluster, seperti prioritas penambahan tenaga pendidik, penguatan kapasitas institusi, dan pemerataan akses pendidikan tinggi berbasis data.

2. BAHAN DAN METODE

2.1. Bahan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan teknik data mining untuk mengelompokkan provinsi di Indonesia berdasarkan karakteristik distribusi perguruan tinggi dan sumber daya pendidikan tinggi. K-Means merupakan algoritma partisi yang secara iteratif menempatkan setiap objek pada centroid terdekat dan memperbarui centroid untuk meminimalkan jumlah kuadrat dalam setiap cluster [11]. Implementasi dilakukan melalui Orange Data Mining, yaitu platform analitik visual berbasis widget yang mendukung pra-pemrosesan, pemodelan, evaluasi, dan visualisasi alur data secara reproduktif.

Data penelitian merupakan data sekunder tahun 2025 yang mencakup 38 provinsi di Indonesia. Variabel yang digunakan terdiri atas jumlah perguruan tinggi, jumlah tenaga pendidik, jumlah mahasiswa, rasio mahasiswa terhadap tenaga pendidik, dan kebutuhan tenaga pendidik. Variabel jumlah perguruan tinggi, jumlah tenaga pendidik, jumlah mahasiswa, dan rasio mahasiswa terhadap tenaga pendidik digunakan sebagai atribut utama dalam proses clustering, sedangkan kebutuhan tenaga pendidik digunakan untuk mendukung interpretasi hasil cluster.

Tahapan penelitian meliputi pengumpulan data, pemeriksaan nilai hilang dan duplikasi, penyesuaian tipe data, analisis distribusi, normalisasi, pemodelan K-Means, evaluasi cluster, dan interpretasi hasil. Empat atribut numerik jumlah perguruan tinggi, tenaga pendidik, mahasiswa, dan rasio mahasiswa-tenaga pendidik distandardisasi menggunakan z-score agar setiap atribut memiliki rerata 0 dan simpangan baku 1. Standardisasi

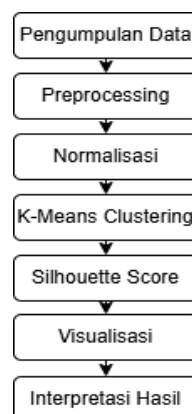
diperlukan karena K-Means berbasis jarak sensitif terhadap perbedaan skala; tanpa normalisasi, jumlah mahasiswa yang bernilai ratusan ribu dapat mendominasi atribut lain [12]. Variabel kebutuhan tenaga pendidik dipertahankan sebagai atribut kategorikal untuk membantu interpretasi hasil dan tidak dimasukkan dalam perhitungan jarak.

Pemodelan dilakukan menggunakan algoritma K-Means dengan tiga cluster ($K=3$) untuk membentuk tipologi kapasitas rendah, menengah atau tertekan, dan tinggi. Pemilihan $K=3$ didasarkan pada kebutuhan interpretasi kebijakan dan dikonfirmasi melalui evaluasi Silhouette. Setiap provinsi dialokasikan ke centroid terdekat menggunakan Euclidean Distance, kemudian centroid diperbarui sampai tidak terjadi perubahan keanggotaan yang berarti. Karena K-Means dapat dipengaruhi centroid awal, proses dijalankan dengan inisialisasi berulang pada Orange Data Mining dan solusi dengan fungsi objektif terbaik digunakan untuk analisis [13], ditunjukkan pada Persamaan 1.

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^n (x_{ik} - c_{jk})^2} \quad (1)$$

$d(x_i, c_j)$ merupakan jarak data ke- i terhadap centroid ke- j , x_{ik} adalah nilai atribut ke- k pada data ke- i , c_{jk} adalah nilai centroid ke- j pada atribut ke- k , dan n adalah jumlah atribut yang digunakan dalam proses clustering.

Evaluasi kualitas cluster dilakukan menggunakan Silhouette Score, yang membandingkan kohesi suatu objek terhadap cluster sendiri dengan separasinya dari cluster terdekat. Nilai mendekati 1 menunjukkan objek sesuai dengan clusternya, nilai sekitar 0 menunjukkan posisi pada batas cluster, dan nilai negatif mengindikasikan potensi salah pengelompokan [14]. Berdasarkan data terstandarisasi dan keanggotaan cluster pada penelitian ini, silhouette rata-rata sebesar 0,520. Hasil clustering kemudian divisualisasikan menggunakan Data Table, Scatter Plot, Distribution, dan Silhouette Plot pada Orange Data Mining untuk mendukung interpretasi.



Gambar 1. Research Methodology

2.2. Studi Literatur

Educational data mining menyediakan pendekatan untuk menemukan pola tersembunyi dalam data pendidikan dan mengubahnya menjadi informasi yang mendukung pengambilan keputusan akademik. Penelitian mutakhir menunjukkan bahwa clustering sering digunakan ketika label kelas tidak tersedia, karena mampu membangun kelompok berdasarkan kemiripan multidimensi dan dapat dikombinasikan dengan klasifikasi maupun association rule mining [15], [16], [17]. Dalam konteks pendidikan tinggi, pendekatan ini telah digunakan untuk mengidentifikasi perilaku, keterlibatan, kinerja, dan risiko pada mahasiswa [18]. Namun, dominasi unit analisis individu membuat pemanfaatannya untuk memetakan kapasitas sistem pendidikan tinggi antarwilayah masih terbatas.

Penelitian Mohamed Nafuri et al. menunjukkan bahwa K-Means dapat digunakan untuk mengelompokkan performa akademik mahasiswa, sedangkan Mohd Talib et al. mengombinasikan K-Means dan Support Vector Machine untuk mengenali pola perilaku mahasiswa. Kajian Le Quy et al. selanjutnya menegaskan bahwa clustering dapat dikombinasikan dengan teknik klasifikasi dan association rule mining untuk memperluas kemampuan analisis dalam educational data mining. Meskipun demikian, ketiga penelitian tersebut masih berfokus pada level mahasiswa dan belum memetakan ketimpangan kapasitas pendidikan tinggi pada tingkat provinsi sebagaimana dilakukan dalam penelitian ini.

K-Means dipilih karena efisien untuk data numerik, mudah direplikasi, dan menghasilkan centroid yang dapat diinterpretasikan sebagai profil rata-rata kelompok. Penerapan clustering pada profil kognitif dan pengelolaan pendidikan juga menunjukkan bahwa segmentasi dapat membantu membedakan karakteristik kelompok untuk mendukung intervensi yang lebih terarah [19], [20], [21]. Kelebihan tersebut sesuai dengan

kebutuhan untuk membangun tipologi provinsi. Meski demikian, K-Means mengasumsikan cluster relatif kompak, sensitif terhadap skala dan penciran, serta memerlukan jumlah cluster sejak awal. Oleh sebab itu, penelitian ini menerapkan standardisasi z-score, pemeriksaan distribusi data, inisialisasi berulang, dan Silhouette Score untuk mengurangi risiko interpretasi yang lemah. Orange Data Mining mendukung analisis melalui antarmuka visual yang menghubungkan widget data, preprocessing, clustering, dan evaluasi dalam satu workflow. Transparansi alur ini memudahkan reproduksi analisis dan pemeriksaan hubungan antara data masukan, hasil cluster, dan visualisasi. Akan tetapi, penggunaan perangkat lunak tidak menggantikan kebutuhan akan desain metodologis yang tepat; pemilihan atribut, normalisasi, nilai K, dan evaluasi tetap menentukan validitas hasil.

Berdasarkan literatur tersebut, penelitian ini memosisikan clustering bukan semata-mata sebagai teknik segmentasi, melainkan sebagai alat penyusunan tipologi kebijakan. Pendekatan pengelompokan siswa yang menghubungkan hasil analitik dengan praktik pedagogis juga mendukung pentingnya menerjemahkan cluster menjadi rekomendasi operasional [22]. Integrasi jumlah institusi, tenaga pendidik, mahasiswa, rasio layanan, dan kebutuhan tenaga pendidik memungkinkan cluster dibaca sebagai profil kapasitas sistem pendidikan tinggi. Kerangka ini berbeda dari penelitian yang hanya menilai performa mahasiswa atau indikator pendidikan tunggal, serta memberikan dasar untuk merancang intervensi yang spesifik bagi kelompok provinsi.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data pendidikan tinggi dari 38 provinsi di Indonesia. Pemeriksaan awal menunjukkan bahwa seluruh provinsi memiliki data lengkap pada atribut utama dan tidak ditemukan duplikasi pada unit analisis. Distribusi data sangat beragam: jumlah perguruan tinggi 7–348, tenaga pendidik 98–24.179, mahasiswa 3.347–671.153, dan rasio mahasiswa-tenaga pendidik 19–40. Perbedaan rentang tersebut menegaskan perlunya standardisasi z-score sebelum melakukan clustering. Setelah normalisasi, seluruh atribut memiliki skala yang sebanding sehingga pembentukan cluster tidak didominasi oleh jumlah mahasiswa atau tenaga pendidik. Distribusi perguruan tinggi dan sumber daya pendidikan tinggi di Indonesia dapat ditunjukkan pada Tabel 1.

Table 1. Distribution of Higher Education Institutions and Resources in Indonesia

Provinsi	Jumlah Perguruan Tinggi	Jumlah Tenaga Pendidik	Jumlah Mahasiswa	Rasio 1:30	Kebutuhan akan Tenaga Pendidik
Aceh	70	3251	80555	25	Tenaga Pendidik Cukup
Sumatera Utara	188	11351	336200	30	Tenaga Pendidik Normal
Sumatera Barat	80	3884	98557	25	Tenaga Pendidik Cukup
Riau	63	3740	116523	31	Tenaga Pendidik Normal
Jambi	33	1485	33833	23	Tenaga Pendidik Cukup
Sumatera Selatan	78	3991	90677	23	Tenaga Pendidik Cukup
Bengkulu	14	1011	27818	28	Tenaga Pendidik Cukup
Lampung	56	2881	75770	26	Tenaga Pendidik Cukup
Kepulauan Bangka Belitung	10	410	7983	19	Tenaga Pendidik Cukup
Kepulauan Riau	26	1453	34539	24	Tenaga Pendidik Cukup
DKI Jakarta	237	24179	666502	28	Tenaga Pendidik Cukup
Jawa Barat	348	22112	671153	30	Tenaga Pendidik Normal
Jawa Tengah	228	14454	485361	34	Tenaga Pendidik Kurang
DI Yogyakarta	101	7915	267524	34	Tenaga Pendidik Kurang
Jawa Timur	318	21090	591212	28	Tenaga Pendidik Cukup
Banten	94	7495	226927	30	Tenaga Pendidik Normal
Bali	51	3832	87638	23	Tenaga Pendidik Cukup
Nusa Tenggara Barat	56	4019	108960	27	Tenaga Pendidik Cukup
Nusa Tenggara Timur	57	2817	92512	33	Tenaga Pendidik Kurang
Kalimantan Barat	41	1579	44831	28	Tenaga Pendidik Cukup
Kalimantan Tengah	25	878	19519	22	Tenaga Pendidik Cukup
Kalimantan Selatan	47	2198	56505	26	Tenaga Pendidik Cukup
Kalimantan Timur	43	1946	52416	27	Tenaga Pendidik Cukup
Kalimantan Utara	7	202	4272	21	Tenaga Pendidik Cukup
Sulawesi Utara	43	1386	36078	26	Tenaga Pendidik Cukup
Sulawesi Tengah	32	1456	40235	28	Tenaga Pendidik Cukup
Sulawesi Selatan	168	10282	262893	26	Tenaga Pendidik Cukup
Sulawesi Tenggara	47	2177	54396	25	Tenaga Pendidik Cukup
Gorontalo	12	1019	22170	22	Tenaga Pendidik Cukup
Sulawesi Barat	17	756	17578	23	Tenaga Pendidik Cukup
Maluku	26	1018	26517	26	Tenaga Pendidik Cukup
Maluku Utara	20	947	19793	21	Tenaga Pendidik Cukup

Provinsi	Jumlah Perguruan Tinggi	Jumlah Tenaga Pendidik	Jumlah Mahasiswa	Rasio 1:30	Kebutuhan akan Tenaga Pendidik
Papua Barat	10	226	8440	37	Tenaga Pendidik Kurang
Papua Barat Daya	12	720	24359	34	Tenaga Pendidik Kurang
Papua	27	905	25393	28	Tenaga Pendidik Cukup
Papua Selatan	7	98	3347	34	Tenaga Pendidik Kurang
Papua Tengah	13	265	6185	23	Tenaga Pendidik Cukup
Papua Pegunungan	8	210	8302	40	Tenaga Pendidik Kurang

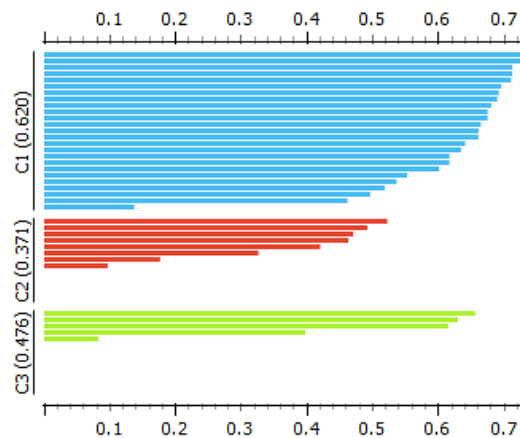
Setelah dilakukan tahap pra-pemrosesan dan normalisasi data, proses pengelompokan dilakukan menggunakan algoritma K-Means dengan jumlah cluster sebanyak 3 kelompok ($K=3$). Pemilihan tiga cluster bertujuan untuk mengidentifikasi karakteristik distribusi perguruan tinggi dan sumber daya pendidikan tinggi berdasarkan tingkat kemiripan setiap provinsi.

Proses clustering menghasilkan tiga kelompok dengan karakteristik yang berbeda. Cluster 1 beranggotakan 25 provinsi dan terbentuk karena nilai jumlah institusi, tenaga pendidik, dan mahasiswa umumnya berada di bawah pusat-pusat pendidikan tinggi nasional, dengan rasio layanan yang relatif moderat. Cluster 2 terdiri atas 8 provinsi yang cenderung memiliki rasio mahasiswa-tenaga pendidik tinggi atau kebutuhan tenaga pendidik yang lebih menonjol, sehingga mencerminkan tekanan pada kapasitas akademik. Cluster 3 beranggotakan 5 provinsi Sumatera Utara, DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur yang memiliki skala institusi, tenaga pendidik, dan mahasiswa jauh lebih besar. Dengan demikian, pembentukan cluster terutama dijelaskan oleh kombinasi skala sistem pendidikan tinggi dan keseimbangan kapasitas tenaga pendidik. Hasil cluster dataset ditunjukkan pada tabel 2.

Table 2. Clustering Result Dataset

Provinsi	Cluster
Aceh	C1
Sumatera Utara	C3
Sumatera Barat	C1
Riau	C2
Jambi	C1
Sumatera Selatan	C1
Bengkulu	C1
Lampung	C1
Kepulauan Bangka Belitung	C1
Kepulauan Riau	C1
DKI Jakarta	C3
Jawa Barat	C3
Jawa Tengah	C3
DI Yogyakarta	C2
Jawa Timur	C3
Banten	C2
Bali	C1
Nusa Tenggara Barat	C1
Nusa Tenggara Timur	C2
Kalimantan Barat	C1
Kalimantan Tengah	C1
Kalimantan Selatan	C1
Kalimantan Timur	C1
Kalimantan Utara	C1
Sulawesi Utara	C1
Sulawesi Tengah	C1
Sulawesi Selatan	C1
Sulawesi Tenggara	C1
Gorontalo	C1
Sulawesi Barat	C1
Maluku	C1
Maluku Utara	C1
Papua Barat	C2
Papua Barat Daya	C2
Papua	C1
Papua Selatan	C2
Papua Tengah	C1
Papua Pegunungan	C2

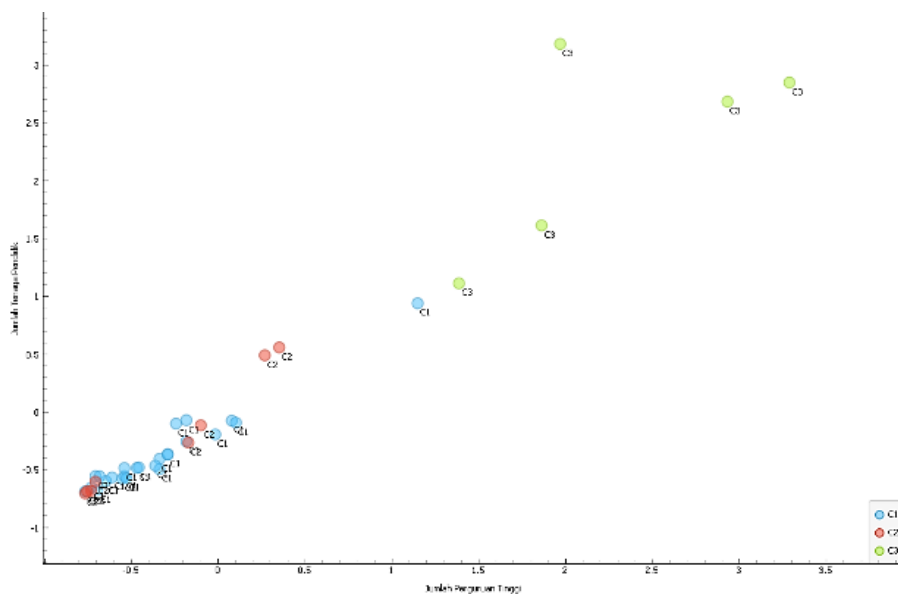
Pengelompokan tersebut menunjukkan bahwa setiap cluster memiliki karakteristik yang berbeda berdasarkan jumlah perguruan tinggi, jumlah tenaga pendidik, jumlah mahasiswa, rasio mahasiswa terhadap tenaga pendidik, dan kebutuhan tenaga pendidik. Perbedaan tersebut menjadi dasar dalam menganalisis kondisi distribusi pendidikan tinggi di Indonesia. Setelah proses clustering dilakukan, kualitas hasil pengelompokan dievaluasi menggunakan Silhouette Plot. Metode ini digunakan untuk mengukur tingkat kemiripan suatu data terhadap cluster tempat data tersebut berada dibandingkan dengan cluster lainnya, Silhouette Plot ditunjukkan pada Gambar 2.



Gambar 2. Silhouette Plot

Evaluasi menghasilkan Silhouette Score rata-rata sebesar 0,520. Nilai positif ini menunjukkan bahwa secara umum anggota lebih dekat dengan cluster sendiri daripada cluster lain. Cluster 1 memiliki pemisahan paling konsisten karena anggotanya relatif homogen, sedangkan beberapa anggota Cluster 2 berada lebih dekat dengan batas cluster akibat kombinasi rasio layanan dan skala institusi yang beragam. Tidak dominannya nilai negatif menunjukkan bahwa struktur tiga cluster cukup layak digunakan untuk interpretasi kebijakan, meskipun belum mencerminkan pemisahan yang sangat kuat.

Nilai silhouette yang diperoleh mengindikasikan bahwa algoritma K-Means mampu menghasilkan pengelompokan yang stabil dan representatif terhadap karakteristik distribusi perguruan tinggi dan sumber daya pendidikan tinggi di Indonesia. Dengan demikian, hasil clustering dapat digunakan sebagai dasar dalam proses analisis dan interpretasi data. Visualisasi hasil clustering dilakukan menggunakan Scatter Plot pada Orange Data Mining untuk melihat penyebaran setiap provinsi berdasarkan karakteristik yang dimiliki, Scatter Plot ditunjukkan pada Gambar 3.



Gambar 3. Scatter Plot

Hasil visualisasi menunjukkan bahwa setiap cluster membentuk kelompok yang relatif terpisah. Cluster pertama memiliki jumlah anggota paling banyak dengan sebaran nilai yang cenderung homogen. Cluster kedua

menunjukkan karakteristik yang berbeda dibandingkan dengan kelompok lainnya, sedangkan Cluster ketiga terdiri atas beberapa provinsi yang memiliki nilai jumlah perguruan tinggi, tenaga pendidik, dan jumlah mahasiswa yang lebih tinggi. Visualisasi tersebut memperlihatkan bahwa algoritma K-Means mampu mengidentifikasi pola distribusi pendidikan tinggi berdasarkan kemiripan karakteristik setiap provinsi sehingga mempermudah proses interpretasi hasil pengelompokan.

Hasil clustering menunjukkan bahwa distribusi perguruan tinggi dan sumber daya pendidikan tinggi di Indonesia masih memiliki karakteristik yang beragam antarprovinsi. Sebagian besar provinsi tergabung dalam Cluster 1, yang menunjukkan kapasitas pendidikan tinggi relatif lebih rendah, dengan distribusi sumber daya yang cukup merata. Cluster 2 memperlihatkan adanya provinsi dengan karakteristik khusus yang memerlukan perhatian terhadap pemenuhan kebutuhan tenaga pendidik. Meskipun memiliki jumlah perguruan tinggi yang cukup, beberapa provinsi dalam kelompok ini masih menunjukkan kebutuhan akan peningkatan kapasitas sumber daya akademik untuk menjaga kualitas layanan pendidikan tinggi. Sementara itu, Cluster 3 terdiri atas provinsi dengan kapasitas pendidikan tinggi yang tinggi dan menjadi pusat aktivitas akademik nasional. Jumlah perguruan tinggi, tenaga pendidik, dan mahasiswa pada kelompok ini relatif lebih besar dibandingkan cluster lainnya sehingga menunjukkan adanya konsentrasi sumber daya pendidikan tinggi pada wilayah tertentu.

Temuan tersebut menunjukkan bahwa pemerataan pendidikan tinggi tidak hanya dipengaruhi oleh jumlah institusi yang tersedia, tetapi juga oleh distribusi tenaga pendidik serta keseimbangan antara jumlah mahasiswa dan kapasitas layanan akademik. Oleh karena itu, hasil clustering dapat dimanfaatkan sebagai dasar penyusunan kebijakan pemerataan pendidikan tinggi, distribusi tenaga pendidik, serta pengembangan kapasitas perguruan tinggi sesuai dengan karakteristik masing-masing wilayah. Hasil penelitian ini memperluas temuan Putri et al. [2] dan Mustakim dan Kamal [3] yang menunjukkan efektivitas K-Means dalam mengelompokkan indikator pendidikan berdasarkan wilayah. Perbedaannya, penelitian ini tidak hanya menggunakan indikator hasil pendidikan, tetapi juga menggabungkan skala institusi, tenaga pendidik, jumlah mahasiswa, rasio layanan, dan kebutuhan tenaga pendidik. Integrasi tersebut menghasilkan tipologi kapasitas pendidikan tinggi yang lebih langsung terkait dengan kebutuhan kebijakan.

Konsentrasi di DKI Jakarta dan provinsi-provinsi besar di Jawa pada Cluster 3 konsisten dengan logika aglomerasi sumber daya pendidikan tinggi: wilayah dengan populasi, pusat pemerintahan, pasar kerja, dan ekosistem akademik yang besar cenderung mengakumulasi institusi, pendidik, dan mahasiswa. Sebaliknya, provinsi baru dan wilayah berpenduduk yang lebih banyak tersebar di Cluster 1 atau Cluster 2 karena skala institusi dan tenaga pendidiknya rendah atau rasio layanannya lebih tertekan. Perbedaan ini bukan semata-mata menunjukkan mutu yang rendah, melainkan menggambarkan kapasitas dan tekanan layanan yang berbeda. Dibandingkan dengan penelitian Talib et al. [6] dan Mohamed Nafuri et al. [8] yang mengelompokkan karakteristik mahasiswa, penelitian ini menunjukkan bahwa teori clustering pendidikan juga dapat diterapkan pada tingkat sistem. Centroid cluster berfungsi sebagai profil kebijakan: Cluster 1 memerlukan penguatan akses dan kapasitas institusi secara selektif; Cluster 2 memerlukan prioritas rekrutmen, redistribusi, atau peningkatan produktivitas tenaga pendidik; sedangkan Cluster 3 memerlukan pengelolaan kepadatan, mutu, serta perannya sebagai pusat dukungan akademik nasional.

Nilai silhouette 0,520 menunjukkan pemisahan cluster yang cukup baik, tetapi juga menegaskan bahwa kondisi pendidikan tinggi antarprovinsi bersifat kontinu dan tidak sepenuhnya terbagi ke dalam kelompok yang kaku. Oleh karena itu, hasil clustering sebaiknya digunakan sebagai alat penyaringan awal, bukan sebagai dasar keputusan yang satu-satunya. Kebijakan perlu dilengkapi dengan indikator mutu, pembiayaan, program studi, kondisi geografis, pendidikan jarak jauh, dan tren data lintas tahun agar intervensi tidak hanya berorientasi pada jumlah institusi.

4. KESIMPULAN

Algoritma K-Means Clustering mampu mengelompokkan distribusi perguruan tinggi dan sumber daya pendidikan tinggi di Indonesia berdasarkan lima indikator, yaitu jumlah perguruan tinggi, jumlah tenaga pendidik, jumlah mahasiswa, rasio mahasiswa terhadap tenaga pendidik, dan kebutuhan tenaga pendidik. Hasil pengelompokan menghasilkan tiga cluster dengan karakteristik yang berbeda, sedangkan evaluasi menggunakan Silhouette Plot menunjukkan bahwa kualitas cluster berada pada kategori baik sehingga mampu merepresentasikan pola distribusi pendidikan tinggi secara memadai.

Hasil penelitian menunjukkan bahwa sebagian besar provinsi berada pada cluster dengan kapasitas pendidikan tinggi relatif rendah, sementara sebagian kecil provinsi memiliki konsentrasi perguruan tinggi, tenaga pendidik, dan mahasiswa yang lebih tinggi. Temuan ini mengindikasikan bahwa distribusi sumber daya pendidikan tinggi di Indonesia masih belum merata, sehingga diperlukan kebijakan yang disesuaikan dengan karakteristik masing-masing wilayah.

Hasil clustering dapat dimanfaatkan sebagai informasi pendukung dalam penyusunan kebijakan pemerataan perguruan tinggi, distribusi tenaga pendidik, serta pengembangan kapasitas pendidikan tinggi berbasis data. Penelitian ini terbatas pada satu periode pengamatan, empat atribut numerik utama, dan penggunaan K-Means yang sensitif terhadap skala, pencilan, serta penentuan jumlah cluster. Variabel kebutuhan tenaga pendidik juga digunakan terutama untuk interpretasi, belum sebagai ukuran numerik dalam

pembentukan cluster. Penelitian selanjutnya perlu menggunakan data longitudinal, menambahkan indikator mutu, pembiayaan, dan kondisi geografis, serta membandingkan K-Means dengan hierarchical clustering, Gaussian mixture model, DBSCAN, atau metode lain. Stabilitas cluster juga perlu diuji melalui beberapa nilai K dan validasi internal maupun eksternal.

REFERENCES

- [1] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, “‘Everything is data’: towards one big data ecosystem using multiple sources of data on higher education in Indonesia,” *J. Big Data*, vol. 9, no. 1, p. 91, 2022, doi: 10.1186/s40537-022-00639-7
- [2] M. R. Putri, G. S. Nugraha, and R. Dwiyanaputra, “Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means Clustering,” *Journal of Computer Science and Informatics Engineering (J-Cosine)*, vol. 7, no. 1, pp. 76–83, 2023, doi: 10.29303/jcosine.v7i1.509
- [3] Z. Mustakim and R. Kamal, “K-means clustering for classifying the quality management of secondary education in Indonesia,” *Cakrawala Pendidikan*, vol. 40, no. 3, pp. 723–735, 2021, doi: 10.21831/cp.v40i3.40150
- [4] C. E. Widiantoro, S. P. Putri, T. Purwandari, and others, “Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Jenjang SMA/Sederajat dengan Analisis Kluster Non Hierarki,” *BIAStatistics Journal Of Statistics Theory And Application Special Issue*, vol. 4, no. 1, pp. 1–9, 2024, doi: 10.1234/biassi.v2024i.273
- [5] U. Arfan and Y. Pekei, “Evaluasi Distribusi Guru-Siswa dan Ketersediaan Sekolah untuk Mendukung Pembangunan Pendidikan Menggunakan K-Means Clustering: An Evaluation of Teacher–Student Distribution and School Availability in Supporting Educational Development Using the K-Means Clustering Algorithm,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 4, pp. 1508–1516, 2025, doi: 10.3390/app13053267.
- [6] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, “Identification of student behavioral patterns in higher education using K-means clustering and support vector machine,” *Applied Sciences*, vol. 13, no. 5, p. 3267, 2023, doi: 10.3390/app13053267.
- [7] L. N. Rani, S. Defit, and L. J. Muhammad, “Determination of student subjects in higher education using hybrid data mining method with the k-means algorithm and fp growth,” *International Journal Of Artificial Intelligence Research*, vol. 5, no. 1, 2021,
- [8] 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,” *Applied Sciences*, vol. 12, no. 19, p. 9467, 2022, doi: 10.3390/app12199467.
- [9] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci. (N. Y.)*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.
- [10] F. Corradini, L. Mozzoni, M. Piangerelli, B. Re, and L. Rossi, “A Framework for Rapidly Prototyping Data Mining Pipelines,” *Big Data and Cognitive Computing*, vol. 9, no. 6, p. 150, 2025, doi: 10.3390/bdcc9060150.
- [11] R. Liu, “Data analysis of educational evaluation using K-means clustering method,” *Comput. Intell. Neurosci.*, vol. 2022, no. 1, p. 3762431, 2022, doi: 10.1155/2022/3762431.
- [12] J. Newling and F. Fleuret, “Fast k-means with accurate bounds,” in *International Conference on Machine Learning*, 2016, pp. 936–944.
- [13] R. Mussabayev, N. Mladenovic, B. Jarboui, and R. Mussabayev, “How to use K-means for big data clustering?,” *Pattern Recognit.*, vol. 137, p. 109269, 2023, doi: 10.1016/j.patcog.2022.109269.
- [14] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
- [15] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student performance prediction using machine learning techniques,” *Educ. Sci. (Basel)*, vol. 11, no. 9, p. 552, 2021, doi: 10.3390/educsci11090552.
- [16] M. A. Al-Barrak and M. Al-Razgan, “Predicting students final GPA using decision trees: a case study,” *International journal of information and education technology*, vol. 6, no. 7, p. 528, 2016, doi: 10.7763/IJiet.2016.V6.745.
- [17] S. M. Dol and P. M. Jawandhiya, “Classification technique and its combination with clustering and association rule mining in educational data mining A survey,” *Eng. Appl. Artif. Intell.*, vol. 122, p. 106071, 2023, doi: 10.1016/j.engappai.2023.106071.
- [18] S. Kim, S. Cho, J. Y. Kim, and D.-J. Kim, “Statistical assessment on student engagement in asynchronous online learning using the k-means clustering algorithm,” *Sustainability*, vol. 15, no. 3, p. 2049, 2023, doi: 10.3390/su15032049.
- [19] M. Orsoni *et al.*, “Preliminary evidence on machine learning approaches for clusterizing students’ cognitive profile,” *Heliyon*, vol. 9, no. 3, 2023, doi: 10.1016/j.heliyon.2023.e14506.

- [20] B. Peng and X. Pei, "A decision support system model for middle school education management based on sparse clustering algorithm," *Mobile Information Systems*, vol. 2022, no. 1, p. 4807395, 2022, doi: 10.1155/2022/4807395.
- [21] Z. Wang, "Higher education management and student achievement assessment method based on clustering algorithm," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, p. 4703975, 2022, doi: 10.1155/2022/4703975.
- [22] C. G. Lechuga and S. Doroudi, "Three algorithms for grouping students: A bridge between personalized tutoring system data and classroom pedagogy," *Int. J. Artif. Intell. Educ.*, vol. 33, no. 4, pp. 843–884, 2023, doi: 10.1007/s40593-022-00309-y.