

# Evaluasi Performa Rmixmod dan KAMILA dalam Pengelompokan Perguruan Tinggi di Indonesia Berdasarkan Data Capaian Kinerja Bertipe Campuran

Andrianto Santoso\*, Anang Kurnia, Aji Hamim Wigena

Sekolah Sains Data, Matematika, dan Informatika, Program Magister Statistika dan Sains Data, IPB University, Bogor, Indonesia

Email: <sup>1,\*</sup>andrianto.santoso@apps.ipb.ac.id, <sup>2</sup>anangk@apps.ipb.ac.id, <sup>3</sup>aji\_hw@apps.ipb.ac.id

Email Penulis Korespondensi: andrianto.santoso@apps.ipb.ac.id

Submitted: 18/05/2025; Accepted: 13/06/2025; Published: 13/06/2025

**Abstrak**—Klasterisasi adalah teknik untuk mengelompokkan objek berdasarkan kesamaan objek di dalam kelompok dan perbedaan di luar kelompok. Dalam dunia nyata, objek sering kali memiliki karakteristik yang diwakili oleh kombinasi peubah numerik dan kategorik, sehingga membutuhkan teknik pengelompokan yang dapat memproses data bertipe campuran. Pengelompokan berbasis model merupakan salah satu pendekatan yang dapat digunakan untuk data tersebut. Penelitian ini mengevaluasi dan membandingkan dua algoritma pengelompokan berbasis model untuk data bertipe campuran, yaitu Rmixmod, yang menggunakan *mixture model* dengan *maximum likelihood estimation* and *expectation-maximization*, dan KAMILA, yang menggunakan pendekatan semi-parametrik. Kedua algoritma tersebut diimplementasikan untuk mengelompokkan perguruan tinggi di Indonesia berdasarkan capaian kerjanya. Jumlah kelompok yang optimal ditentukan dengan menggunakan *Bayesian Information Criterion* dan Koefisien *Silhouette*. Sedangkan kinerja algoritma dievaluasi dengan menggunakan Koefisien *Silhouette*, Indeks Calinski-Harabasz, dan Indeks Davies-Bouldin. Hasil penelitian menunjukkan bahwa algoritma Rmixmod unggul dibandingkan KAMILA dalam pengelompokan perguruan tinggi di Indonesia, dengan nilai Koefisien *Silhouette* sebesar 0,2878, Indeks Calinski-Harabasz sebesar 253,9433, dan Indeks Davies-Bouldin sebesar 1,5321. Jumlah kelompok optimal yang terbentuk adalah lima kelompok. Interpretasi kelompok dilakukan dengan menganalisis nilai rata-rata PC dan distribusi peubah kategorik dalam setiap kelompok. Hasil pengelompokan ini diharapkan dapat menjadi landasan bagi pemerintah dalam merumuskan kebijakan strategis yang tepat guna dan terdiferensiasi sesuai dengan karakteristik masing-masing kelompok perguruan tinggi.

**Kata Kunci:** Perguruan Tinggi; Pengelompokan Berbasis Model; Data Campuran; Evaluasi Pengelompokan; Pembuatan Kebijakan

**Abstract**—Clustering is a technique for grouping objects based on their similarities within clusters and their differences across clusters. In real-world, objects often have characteristics represented by a combination of numerical and categorical variables, requiring clustering techniques that can process mixed-type data. Model-based clustering is one of the approaches that can be utilized for such data. This study evaluates and compares two model-based clustering algorithms for mixed data type, Rmixmod, which employs a mixture model with maximum likelihood estimation and expectation-maximization, and KAMILA, which utilizes a semi-parametric approach. Both algorithms are implemented to cluster Indonesian higher education institutions based on their performance. The optimal number of clusters is determined using the Bayesian Information Criterion and the Silhouette Coefficient. Algorithms performance is evaluated using the Silhouette Coefficient, the Calinski-Harabasz Index, and the Davies-Bouldin Index. The research results showed that the Rmixmod algorithm outperformed KAMILA in clustering Indonesian higher education institutions, with a Silhouette Coefficient of 0.2878, a Calinski-Harabasz Index of 253.9433, and a Davies-Bouldin Index of 1.5321. The optimal number of clusters formed was five. Cluster interpretation is conducted by analyzing the mean values of PC and the distribution of categorical variables within each cluster. The clustering results are expected to serve as a foundation for the government in formulating strategic policies that are both effective and differentiated according to the characteristics of each group of higher education institutions.

**Keywords:** Higher Education Institutions; Model-Based Clustering; Mixed-Type Data; Clustering Evaluation; Policymaking

## 1. PENDAHULUAN

Capaian kinerja perguruan tinggi merupakan kemampuan institusi dalam melaksanakan tri dharma perguruan tinggi yang meliputi pendidikan, penelitian, dan pengabdian kepada masyarakat. Untuk menilai capaian kinerja tersebut, terdapat beberapa indikator yang dapat diukur, seperti kualitas lulusan (tingkat kelulusan, masa tunggu kerja, kepuasan pengguna), produktivitas penelitian (jumlah publikasi, hak kekayaan intelektual), serta dampak pengabdian kepada masyarakat (program kolaboratif, pemberdayaan lokal, dan inovasi sosial). Selain itu, efektivitas tata kelola perguruan tinggi dalam mengelola sumber daya yang dimiliki berupa sumber daya manusia, sarana dan prasarana, serta keuangan juga menjadi tolok ukur yang penting. Secara eksternal, adanya akreditasi perguruan tinggi oleh Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) dan pemeringkatan internasional (*QS World University Rankings*, *Times Higher Education*) menjadi alat kontrol terhadap kualitas perguruan tinggi [1].

Perguruan tinggi di Indonesia yang berjumlah 4.437 perguruan tinggi [2] menciptakan kompleksitas tersendiri bagi Kementerian Pendidikan Tinggi, Sains, dan Teknologi dan kementerian lainnya yang memiliki perguruan tinggi dalam mendorong peningkatan mutu pendidikan tinggi secara merata. Untuk merumuskan kebijakan strategis yang efektif, diperlukan pendekatan sistematis dalam memetakan heterogenitas kapasitas dan capaian kinerja institusi. Salah satu solusi yang relevan adalah klasterisasi perguruan tinggi.

Klasterisasi (pengelompokan) merupakan teknik untuk mengelompokkan objek berdasarkan tingkat kesamaan atau ketidakmiripan antar objek, sehingga objek dengan karakteristik serupa tergabung dalam satu kelompok yang homogen [3]. Teknik ini digunakan untuk menggali wawasan dari data yang tersedia, yang selanjutnya dapat

dimanfaatkan dalam berbagai kepentingan, seperti pengambilan kebijakan berbasis data [4], [5]. Dalam konteks perguruan tinggi, pengelompokan dapat dilakukan berdasarkan karakteristik dan indikator capaian kinerja terukur, seperti produktivitas penelitian, produktivitas pengabdian kepada masyarakat, akreditasi perguruan tinggi, jumlah dosen berdasarkan tingkatan jabatan akademik, dan indikator lain yang relevan.

Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat melakukan pengelompokan perguruan tinggi tahun 2024 berdasarkan capaian kinerja perguruan tinggi dengan kriteria dan bobot sebagai berikut: kelembagaan 15%, sumber daya manusia 15%, penelitian 15%, pengabdian kepada masyarakat 15%, publikasi 25%, dan kekayaan intelektual 10% [6]. Pengelompokan dilakukan pada perguruan tinggi akademik (universitas, institut, sekolah tinggi) di bawah koordinasi Kementerian Pendidikan, Riset, dan Teknologi. Penelitian [7] mengelompokkan Perguruan Tinggi Keagamaan Islam Negeri (PTKIN) berdasarkan indikator *World Class University* dengan metode K-Means. Sedangkan penelitian [8] mengelompokkan perguruan tinggi swasta di lingkungan Lembaga Layanan Pendidikan Tinggi Wilayah V, Yogyakarta berdasarkan Indikator Kinerja Utama Perguruan Tinggi dengan metode *Quartile Binning Method*, dan penelitian [9] mengelompokkan perguruan tinggi swasta berdasarkan minat siswa dengan metode K-Medoids.

Mengacu penelitian yang telah dilakukan sebelumnya, secara umum pengelompokan perguruan tinggi menggunakan indikator berupa data numerik, namun terdapat atribut pada perguruan tinggi berupa data kategorik seperti akreditasi institusi, status perguruan tinggi, letak geografis (provinsi), dan sebagainya. Hal ini menjadi tantangan tersendiri ketika data numerik dan kategorik digabungkan (data campuran) dan digunakan sebagai parameter dalam melakukan pengelompokan. Sehingga diperlukan algoritma pengelompokan yang dapat menangani data campuran secara langsung tanpa perlu melakukan diskretisasi yakni mengubah data numerik menjadi kategorik [10] ataupun *encoding* di mana data kategorik diubah menjadi numerik [11]. Tantangan utama algoritma dalam menangani data campuran terletak pada penentuan metode yang tepat untuk mengukur tingkat kesamaan atau ketidakmiripan antar objek dalam proses pengelompokan [12].

Algoritma pengelompokan untuk data campuran telah banyak dikembangkan. Algoritma tersebut terdiri dari beberapa jenis seperti algoritma partisi, hirarki, berbasis model, berbasis jaringan saraf, dan lainnya [13]. Algoritma partisi memiliki keunggulan berupa kompleksitas waktu yang linear namun memiliki permasalahan dalam hal inisialisasi pusat kelompok dan penentuan jumlah kelompok yang tepat. Sedangkan pada algoritma selain partisi mampu menghasilkan kelompok yang lebih baik namun memiliki kompleksitas waktu yang tidak linear hingga mencapai  $O(n^3)$ .

Dalam penelitian ini, algoritma pengelompokan berbasis model diterapkan untuk mengelompokkan perguruan tinggi. Secara umum algoritma pengelompokan berbasis model menggunakan pendekatan statistik berupa model probabilistik. Penentuan jumlah kelompok dan metode pengelompokan yang tepat berkaitan dengan pemilihan model probabilitas [14]. Selain itu, pendekatan berbasis model tidak mengharuskan standardisasi data sebelum dilakukan pengelompokan [15]. Algoritma pengelompokan berbasis model yang digunakan adalah Rmixmod dan KAMILA (*KAy-means for Mixed Large data sets*). Rmixmod merupakan metode pengelompokan berbasis *mixture model*, di mana data numerik dimodelkan menggunakan distribusi Gaussian dan data kategorik dimodelkan dengan menggunakan distribusi multinomial [16]. Algoritma ini menggunakan pendekatan *Maximum Likelihood Estimation* (MLE) dan *Expectation-Maximization* (EM) untuk menentukan parameter terbaik dalam model pengelompokan. Disamping itu, Rmixmod dapat diimplementasikan pada dataset berukuran besar [17].

Sementara itu, KAMILA merupakan metode pengelompokan yang secara khusus dikembangkan untuk menangani data campuran dengan pendekatan semi-parametrik [18]. KAMILA menggabungkan estimasi kepadatan kernel (*kernel density estimation*) radial untuk data numerik dan model multinomial untuk data kategorik. Berbeda dengan metode berbasis *mixture model*, KAMILA tidak memerlukan asumsi distribusi yang ketat terhadap data numerik dan kategorik [19]. Selain itu, KAMILA memiliki kecepatan dan efisiensi secara komputasi sehingga dapat diterapkan pada dataset berukuran besar [20].

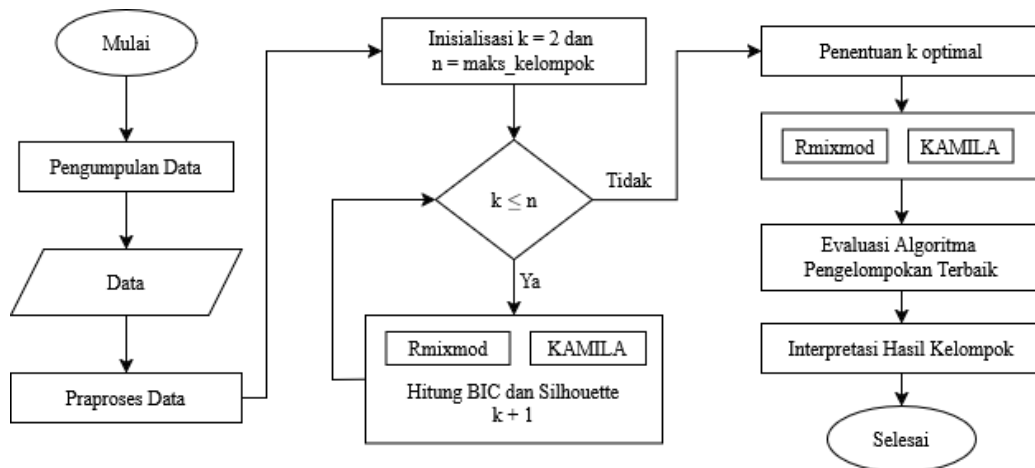
Penelitian ini mengevaluasi performa dari Rmixmod dan KAMILA dalam melakukan pengelompokan perguruan tinggi di Indonesia berdasarkan data capaian kinerja perguruan tinggi yang berupa data campuran. Hasil pengelompokan diharapkan dapat menjadi dasar bagi pemerintah dalam membuat kebijakan strategis yang efektif serta terdiferensiasi sesuai dengan karakteristik tiap kelompok perguruan tinggi.

## 2. METODOLOGI PENELITIAN

Penelitian ini dilakukan melalui tahapan-tahapan mulai dari pengumpulan data, praproses data, penentuan jumlah kelompok optimal, evaluasi algoritma pengelompokan terbaik, dan interpretasi hasil kelompok. Secara keseluruhan, diagram alur penelitian tersaji pada Gambar 1.

Berdasarkan Gambar 1, tahapan penelitian dimulai dengan pengumpulan data dengan teknik *web scraping* (ekstraksi data) dari laman *Science and Technology Index* [21], Badan Akreditasi Nasional Perguruan Tinggi [22], Direktorat Pembinaan Pengelolaan Keuangan BLU [23], dan Pangkalan Data Pendidikan Tinggi [24].

Data yang diperoleh kemudian diproses melalui beberapa langkah praproses, meliputi penggabungan dan filterisasi data, pengolahan dan pembersihan data, serta pengecekan *skewness* dan transformasi pada peubah numerik yang yang menjulur ke kanan. Selanjutnya dilakukan reduksi dimensi guna mengurangi kompleksitas data [25].



Gambar 1. Diagram Alur Penelitian

Pengelompokan dilakukan pada data hasil praproses menggunakan dua algoritma, yaitu Rmixmod dan KAMILA. Proses penentuan jumlah kelompok optimal ( $k$ ) dilakukan dengan iterasi pengujian berbagai nilai  $k$ , dan pemilihan nilai optimal berdasarkan *Bayesian Information Criteria* (BIC) dan koefisien *Silhouette* [16], [20], [26].

Evaluasi kualitas hasil pengelompokan dilakukan menggunakan tiga metrik utama, yaitu koefisien *Silhouette*, indeks Calinski-Harabasz, dan indeks Davies-Bouldin [27]. Ketiga metrik ini digunakan untuk menilai sejauh mana hasil pengelompokan menghasilkan kelompok yang kompak secara internal dan terpisah secara eksternal.

Koefisien *Silhouette* ( $S$ ) mengukur seberapa baik suatu objek cocok dengan kelompok tempat ia berada dibandingkan dengan kelompok terdekat lainnya. Nilai  $S$  maksimum menandakan pengelompokan yang lebih baik. Rata-rata nilai  $S$  dari seluruh objek dihitung menggunakan persamaan (1):

$$S = \frac{1}{n} \sum_{i=1}^n \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (1)$$

dengan  $a_i$  rata-rata jarak antara objek ke- $i$  dan seluruh objek lain dalam kluster yang sama, dan  $b_i$  rata-rata jarak terkecil antara objek ke- $i$  dan objek pada kelompok tetangga terdekat.

Indeks Calinski-Harabasz ( $CH$ ) mengevaluasi rasio rata-rata *sum of squares* antar kelompok dan dalam kelompok. Semakin tinggi nilai  $CH$ , semakin baik pemisahan antar kelompok yang dihasilkan. Nilai  $CH$  dihitung menggunakan persamaan (2):

$$CH = \frac{T_r(B_k)}{T_r(W_k)} \times \frac{N-k}{k-1} \quad (2)$$

dengan  $B_k$  matriks dispersi antar kelompok,  $W_k$  matriks dispersi dalam kelompok,  $N$  jumlah objek, dan  $k$  jumlah kelompok.

Indeks Davies-Bouldin ( $DB$ ) mengukur rata-rata tingkat kemiripan maksimum antar kelompok. Nilai  $DB$  yang lebih rendah menunjukkan kelompok yang lebih baik karena antar kelompok lebih berbeda secara signifikan. Rumus perhitungan  $DB$  diberikan pada persamaan (3):

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left( \frac{S_i + S_j}{d(c_i, c_j)} \right) \quad (3)$$

dengan  $k$  jumlah kelompok,  $c_i$  dan  $c_j$  pusat dari kelompok ke- $i$  dan ke- $j$ ,  $S_i$  dan  $S_j$  jarak rata-rata dari objek dalam kelompok  $i$  dan  $j$  terhadap pusat masing-masing, dan  $d(c_i, c_j)$  jarak antara pusat  $c_i$  dan  $c_j$ .

Ketiga metrik evaluasi tersebut bergantung pada perhitungan jarak antar objek, baik dalam satu kelompok maupun antar kelompok. Untuk data dengan tipe campuran, perhitungan jarak dilakukan menggunakan metode Gower [28]. Setelah hasil pengelompokan diperoleh, masing-masing kelompok dianalisis dan diinterpretasikan berdasarkan rata-rata skor komponen utama serta distribusi peubah kategorik guna mengidentifikasi karakteristik khas tiap kelompok. Selain itu, peubah penciri kelompok dapat dikenali melalui peubah-peubah dominan dalam pembentukan komponen utama [29].

### 3. HASIL DAN PEMBAHASAN

#### 3.1 Pengumpulan Data

Hasil penggabungan dari ekstraksi data diperoleh data sebanyak 4.699 perguruan tinggi yang selanjutnya dilakukan praproses data berupa filterisasi dengan kriteria perguruan tinggi memiliki program studi pascasarjana, pengolahan,

dan pembersihan data sehingga jumlah data menjadi 803 perguruan tinggi dengan peubah numerik sebanyak 33 peubah dan peubah kategorik sebanyak 3 peubah. Rincian peubah disajikan dalam Tabel 1.

**Tabel 1.** Rincian peubah

Peubah	Nama Peubah	Jenis Peubah
X1	Jumlah <i>Author</i>	Numerik
X2	Jumlah Departemen	Numerik
X3	Jumlah Jurnal	Numerik
X4	<i>Sinta Score Overall</i>	Numerik
X5	<i>Sinta Score 3Yr</i>	Numerik
X6	<i>Sinta Score Produktivitas</i>	Numerik
X7	<i>Sinta Score Produktivitas 3Yr</i>	Numerik
X8	Produktivitas Publikasi	Numerik
X9	Produktivitas Publikasi 3Yr	Numerik
X10	Produktivitas Hak Kekayaan Intelektual	Numerik
X11	Produktivitas Hak Kekayaan Intelektual 3Yr	Numerik
X12	Kelembagaan	Numerik
X13	Kelembagaan 3Yr	Numerik
X14	Produktivitas Penelitian	Numerik
X15	Produktivitas Penelitian 3Yr	Numerik
X16	Produktivitas Pengabdian Kepada Masyarakat	Numerik
X17	Produktivitas Pengabdian Kepada Masyarakat 3Yr	Numerik
X18	Total Profesor <i>Overall</i>	Numerik
X19	Total Profesor 3Yr	Numerik
X20	Total Lektor Kepala <i>Overall</i>	Numerik
X21	Total Lektor Kepala 3Yr	Numerik
X22	Total Lektor <i>Overall</i>	Numerik
X23	Total Lektor 3Yr	Numerik
X24	Total Asisten Ahli <i>Overall</i>	Numerik
X25	Total Asisten Ahli 3Yr	Numerik
X26	Total Non-Jafa <i>Overall</i>	Numerik
X27	Total Non-Jafa 3Yr	Numerik
X28	Total Dosen <i>Overall</i>	Numerik
X29	Total Dosen 3Yr	Numerik
X30	Total Prodi S2	Numerik
X31	Total Prodi S2 Terapan	Numerik
X32	Total Prodi S3	Numerik
X33	Total Prodi Pascasarjana	Numerik
X34	Akreditasi Institusi Perguruan Tinggi	Kategorik
X35	Status Perguruan Tinggi	Kategorik
X36	Provinsi	Kategorik

3Yr: 2021-2023

Eksplorasi peubah numerik khususnya untuk melihat sebaran data maka dilakukan perhitungan nilai minimum, nilai rata-rata, nilai maksimum dan nilai *skewness* sebagaimana terdapat pada Tabel 2.

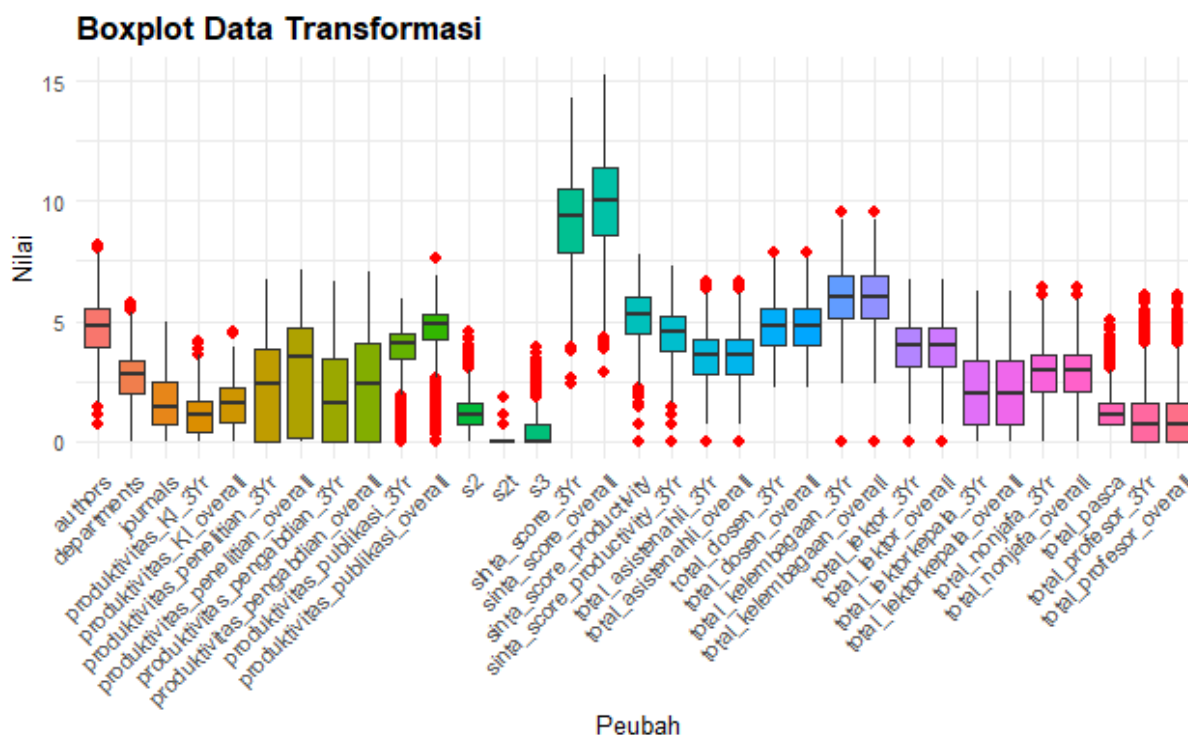
**Tabel 2.** Ringkasan statistik peubah numerik

Peubah	Nilai Minimum	Rata-rata	Nilai Maksimum	Nilai <i>Skewness</i>
X1	1	253	3463	3,44
X2	0	23,67	309	3,88
X3	0	9,164	142	3,33
X4	16	119.751	4.188.574	6,31
X5	10	53.900	1.547.647	5,82
X6	0	290	2.403	2,56
X7	0	139,9	1.444	2,89
X8	0	154,51	2.038,79	4,09
X9	0	68,13	376,78	1,46
X10	0	6,28	91,466	4,37
X11	0	3,37	61,07	5,32
X12	0	923,6	14.470	3,89
X13	0	923,6	14.470	3,89
X14	0	77,13	1.228,37	3,50



Peubah	Nilai Minimum	Rata-rata	Nilai Maksimum	Nilai <i>Skewness</i>
X15	0	37,08	820,41	4,59
X16	0	47,78	1.157,73	4,71
X17	0	26,67	772,14	5,35
X18	0	11,73	428	5,84
X19	0	11,73	428	5,84
X20	0	32,95	524	3,39
X21	0	32,95	524	3,39
X22	0	100,7	812	2,45
X23	0	100,7	812	2,45
X24	0	59,48	742	3,47
X25	0	59,48	742	3,47
X26	0	31,24	623	5,44
X27	0	31,24	623	5,44
X28	8	236,1	2.665	3,06
X29	8	236,1	2.665	3,06
X30	0	4,35	96	5,15
X31	0	0,06	5	8,09
X32	0	1,12	51	7,14
X33	1	5,53	148	5,82

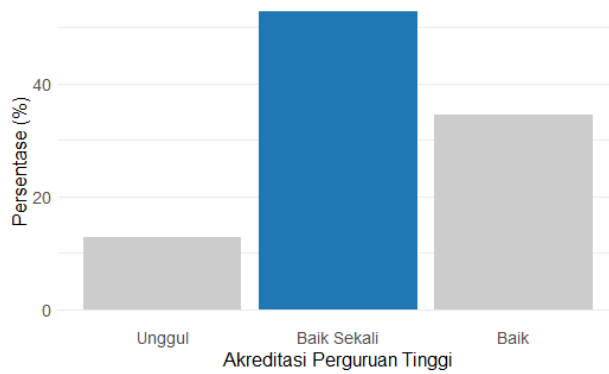
Berdasarkan Tabel 2, selisih nilai maksimum dengan nilai minimum sangat besar pada seluruh peubah numerik. Selain itu, nilai *skewness* pada seluruh peubah numerik bernilai lebih dari 0, hal ini menunjukkan bahwa sebaran data pada setiap peubah numerik menjulur ke kanan. Untuk mengurangi efek *skewness* dan nilai ekstrem maka dilakukan transformasi pada seluruh peubah numerik berupa transformasi log (n+1).



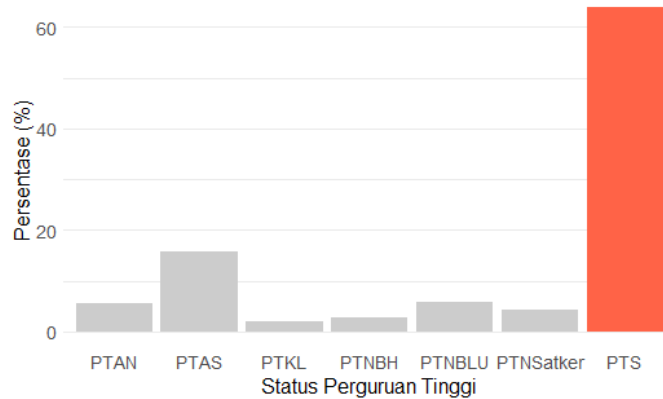
Gambar 2. Boxplot Peubah Numerik Hasil Transformasi

Gambar 2 menunjukkan bahwa distribusi nilai pada peubah numerik menjadi lebih simetris dan merata dibandingkan kondisi sebelum transformasi. Transformasi yang diterapkan berhasil mengurangi efek *skewness*. Meskipun beberapa nilai pencilan masih terdeteksi, nilai-nilai ekstrem tersebut sudah lebih terkendali dan tidak mendominasi distribusi data.

Eksplorasi terhadap peubah kategorik dilakukan dengan menghitung persentase tiap kategori pada masing-masing peubah kategorik. Gambar 3 menunjukkan bahwa akreditasi perguruan tinggi terdiri dari Baik Sekali sebesar 52,68%, diikuti oleh akreditasi Baik sebesar 34,50%, dan Unggul sebesar 12,82%.

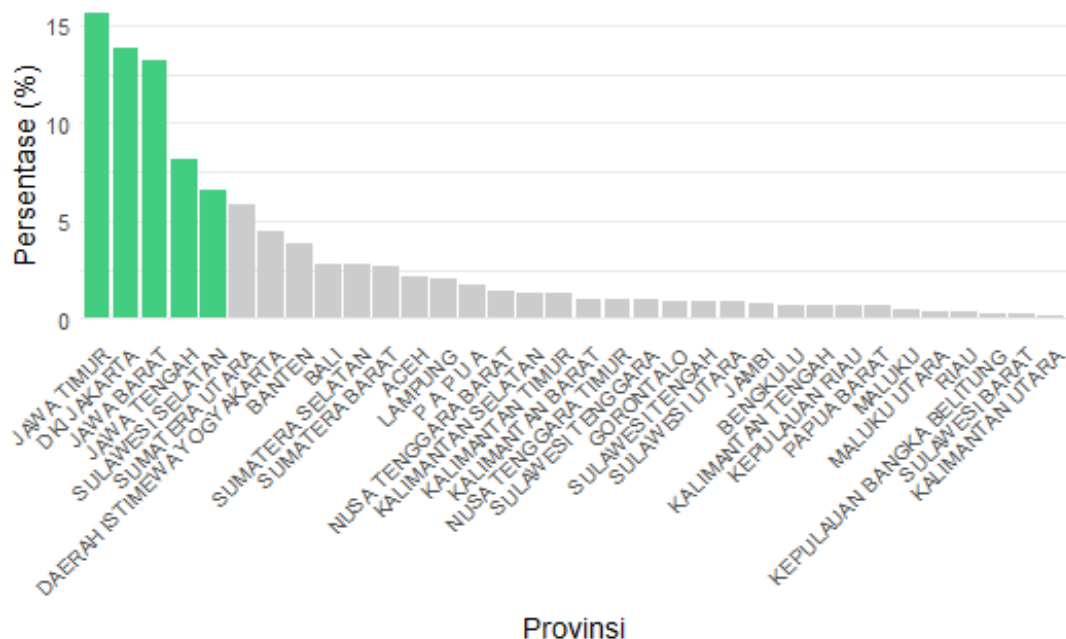


Gambar 3. Barplot Akreditasi Perguruan Tinggi



Gambar 4. Barplot Status Perguruan Tinggi

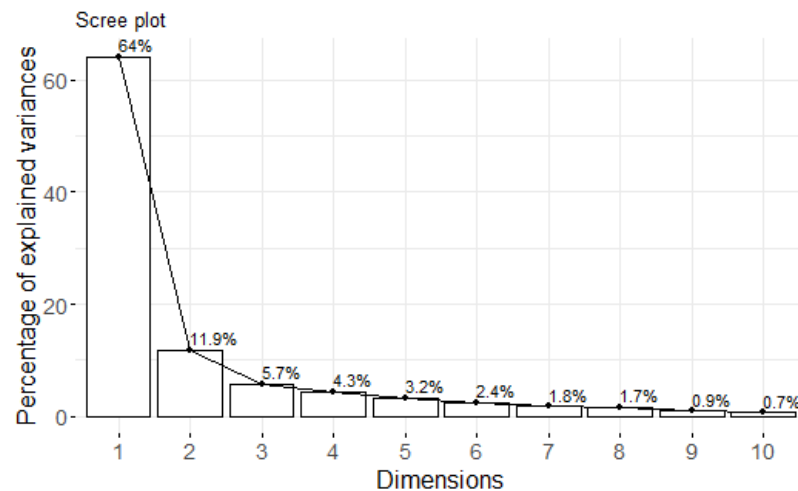
Berdasarkan sebaran status perguruan tinggi, Gambar 4 memperlihatkan Perguruan Tinggi Swasta (PTS) mendominasi sebesar 63,80%, diikuti oleh Perguruan Tinggi Agama Swasta (PTAS) sebesar 15,80%, dan berbagai bentuk Perguruan Tinggi Negeri seperti Perguruan Tinggi Negeri Badan Layanan Umum (PTNBLU) sebesar 5,70%, Perguruan Tinggi Agama Negeri (PTAN) sebesar 5,60%, Perguruan Tinggi Negeri Satuan Kerja (PTNSatker) sebesar 4,20%, Perguruan Tinggi Negeri Berbadan Hukum (PTNBH) sebesar 2,90%, dan Perguruan Tinggi Kementerian Lain atau Lembaga Pemerintah Nonkementerian (PTKL) sebesar 2,00%. Selanjutnya menurut sebaran geografis, provinsi dengan jumlah perguruan tinggi terbanyak adalah Jawa Timur 15,57%, DKI Jakarta 13,82%, Jawa Barat 13,20%, Sulawesi Selatan 6,60%, dan Sumatera Utara 5,85%, sebagaimana dapat dilihat pada Gambar 5.



Gambar 5. Barplot Provinsi

### 3.2 Reduksi Dimensi

Pada penelitian ini karena data sebagian besar berupa peubah numerik maka reduksi dimensi hanya diterapkan pada peubah numerik dengan menggunakan *Principal Component Analysis* (PCA). Penentuan jumlah komponen utama (PC) yang digunakan berdasarkan nilai kumulatif variansi.



**Gambar 6.** Screeplot Persentase Kumulatif Variansi PCA

Berdasarkan Gambar 6, nilai kumulatif variansi sampai dengan PC5 sebesar 89,1%. Sehingga komponen utama yang digunakan yakni PC1 sampai dengan PC5. Pada setiap PC terdapat peubah numerik yang dominan. Hal ini dapat diidentifikasi berdasarkan nilai *loading* masing-masing peubah numerik. Peubah numerik yang paling dominan, ditentukan dari tiga nilai *loading* terbesar untuk setiap PC, disajikan secara rinci dalam Tabel 3.

**Tabel 3.** Peubah numerik dominan

PC	Peubah Numerik Dominan (nilai <i>loading</i> )
1	Sinta <i>Score</i> 3Yr (-0,21), Jumlah <i>Author</i> (-0,21), Total Dosen <i>Overall</i> (-0,20)
2	Sinta <i>Score</i> Produktivitas 3Yr (0,27), Produktivitas Penelitian (0,26), Produktivitas Pengabdian Kepada Masyarakat (0,25)
3	Total Non-Jafa <i>Overall</i> (0,55), Total Asisten Ahli <i>Overall</i> (0,55), Produktivitas Publikasi (0,18)
4	Total Prodi S2 Terapan (0,50), Total Prodi S3 (0,36), Total Prodi S2 (0,34)
5	Total Prodi S2 Terapan (0,61), Produktivitas Publikasi 3Yr (0,31), Produktivitas Pengabdian Kepada Masyarakat 3Yr (0,29)

Berdasarkan Tabel 3, pada PC1 peubah numerik dominan terdiri dari Sinta *Score* 3Yr, Jumlah *Author*, dan Total Dosen *Overall* dengan nilai *loading* masing-masing sebesar -0,21, -0,21, dan -0,20. Hal ini menunjukkan keterkaitan Sinta *Score* dengan jumlah sumber daya dosen. Nilai *loading* negatif mengindikasikan kontribusi peubah dalam arah yang berlawanan terhadap komponen ini. PC2 menampilkan peubah dominan yang berkaitan dengan produktivitas dalam tri dharma perguruan tinggi, yaitu Sinta *Score* Produktivitas 3Yr (0,27), Produktivitas Penelitian (0,26), dan Produktivitas Pengabdian Kepada Masyarakat (0,25).

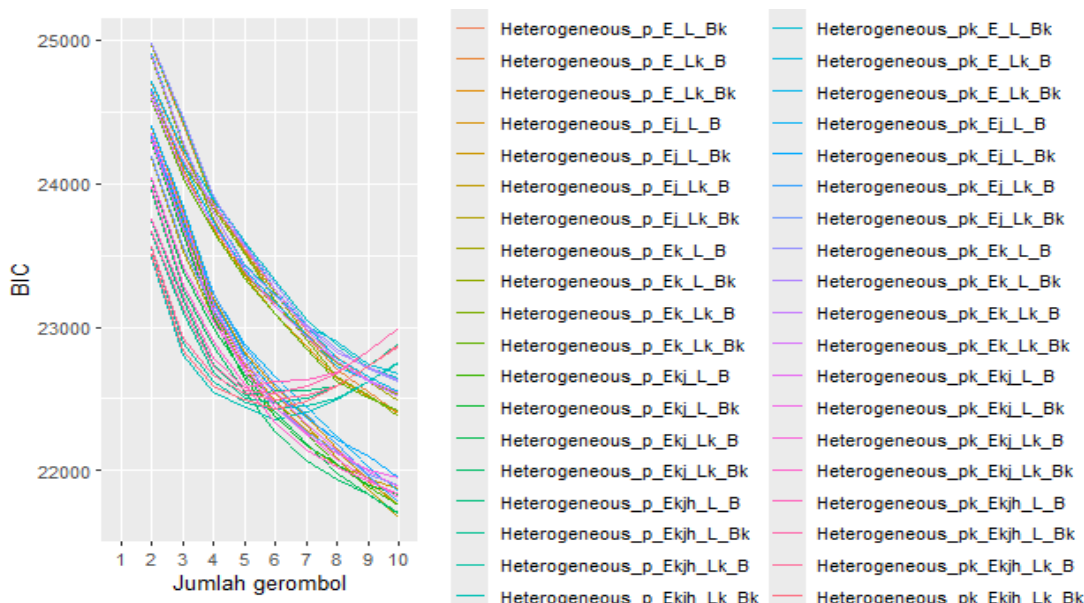
Selanjutnya, PC3 didominasi oleh Total Non-Jafa *Overall*, dan Total Asisten Ahli *Overall* dengan nilai *loading* cukup tinggi sebesar 0,55, serta Produktivitas Publikasi sebesar 0,18, yang menggambarkan fokus komponen ini pada jabatan fungsional dan produktivitas publikasi. Pada PC4, peubah dominan adalah Total Prodi S2 Terapan dengan nilai *loading* 0,50, Total Prodi S3 0,36, dan Total Prodi S2 0,34, yang menunjukkan konsentrasi pada keberagaman program studi pascasarjana. Sedangkan pada PC5, Total Prodi S2 Terapan memiliki nilai *loading* paling tinggi sebesar 0,61, diikuti oleh Produktivitas Publikasi 3Yr (0,31) dan Produktivitas Pengabdian Kepada Masyarakat 3Yr (0,29), yang memperlihatkan keterkaitan erat antara jumlah program studi terapan dan aktivitas produktivitas dalam aspek publikasi serta pengabdian masyarakat.

Hasil dari reduksi dimensi peubah numerik yakni PC1 sampai dengan PC5 digabungkan dengan peubah kategorik sebagai data campuran untuk selanjutnya dilakukan pengelompokan perguruan tinggi dengan menggunakan algoritma Rmixmod dan KAMILA.

### 3.3 Pengelompokan dengan Rmixmod

Pada Rmixmod, pengelompokan diawali dengan penentuan model dan jumlah kelompok (k) yang optimal. Hal ini difasilitasi secara otomatis oleh Rmixmod, namun memerlukan waktu komputasi yang lama. Rmixmod memodelkan data dengan seluruh kemungkinan model untuk data campuran yang berjumlah 40 model dan berbagai nilai k (k = 2,

3, 4, ..., 10). Hasil yang diperoleh yakni model Heterogeneous\_p\_Ej\_Lk\_B dengan nilai k = 10 memiliki nilai BIC paling minimum sebesar 21.671,33 sebagaimana tersaji pada Gambar 7.



Gambar 7. Hasil Pemodelan Data dan Jumlah Kelompok

Selain itu, Gambar 7 juga menunjukkan bahwa pada sebagian besar model, nilai BIC cenderung semakin kecil seiring dengan bertambahnya jumlah kelompok. Oleh karena itu, untuk memastikan jumlah kelompok optimal, digunakan pula koefisien *Silhouette*, di mana kualitas kelompok ditunjukkan oleh nilai koefisien *Silhouette* yang maksimum.

Tabel 4. Nilai BIC dan *Silhouette* hasil pengelompokan Rmixmod

k	BIC	<i>Silhouette</i>
2	24.290,10	0,2718
3	23.683,53	0,2830
4	23.195,24	0,2482
5	22.802,37	0,2878
6	22.536,69	0,1524
7	22.308,65	0,1265
8	22.096,57	0,1367
9	21.858,36	0,1370
10	21.753,77	0,1178

Berdasarkan Tabel 4, jumlah kelompok optimal ditetapkan sebanyak k = 5 karena memenuhi dua kriteria, yaitu nilai BIC minimum dan koefisien *Silhouette* maksimum.

### 3.4 Pengelompokan dengan KAMILA

KAMILA tidak memfasilitasi secara otomatis dalam penentuan jumlah kelompok optimal sehingga dilakukan perhitungan nilai BIC dan koefisien *Silhouette* secara iterasi dengan berbagai nilai k (k = 2, 3, 4, ..., 10).

Tabel 5. Nilai BIC dan *Silhouette* hasil pengelompokan KAMILA

K	BIC	<i>Silhouette</i>
2	18.893,14	0,1501
3	17.223,83	0,2042
4	16.581,83	0,1095
5	15.897,82	0,1262
6	15.374,11	0,0808
7	14.968,87	0,0700
8	14.595,43	0,0573
9	14.126,91	0,0774
10	13.955,81	0,0573

Berdasarkan Tabel 5, jumlah kelompok optimal ditetapkan sebanyak k = 5 karena memenuhi dua kriteria, yaitu nilai BIC minimum dan koefisien *Silhouette* maksimum.

### 3.5 Evaluasi Algoritma pengelompokan Terbaik

Kriteria algoritma pengelompokan terbaik adalah memiliki nilai koefisien *Silhouette* maksimum, indeks Calinski-Harabasz maksimum, dan indeks Davies-Bouldin minimum.

**Tabel 6.** Perbandingan performa Rmixmod dan KAMILA

Metode	<i>Silhouette</i>	Indeks Calinski-Harabasz	Indeks Davies-Bouldin
Rmixmod	0,2878	253,9433	1,5321
KAMILA	0,1262	201,3294	2,1545

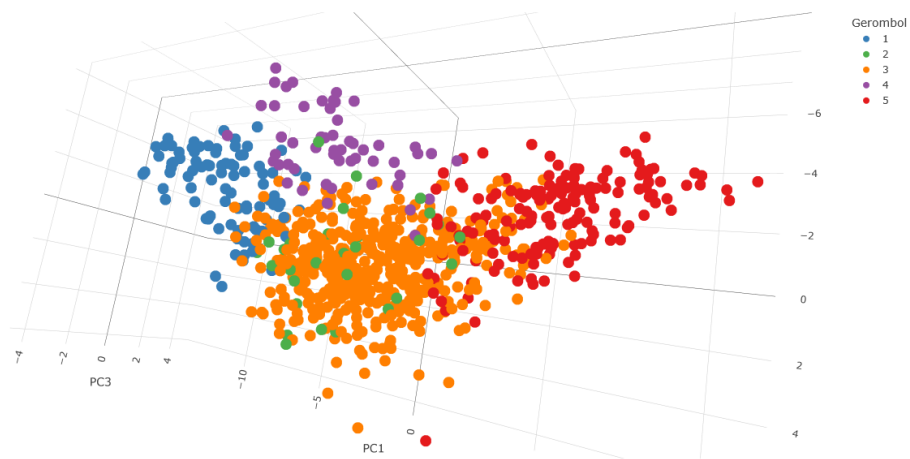
Berdasarkan Tabel 6, algoritma terbaik dalam pengelompokan perguruan tinggi adalah Rmixmod.

### 3.6 Interpretasi Hasil Kelompok

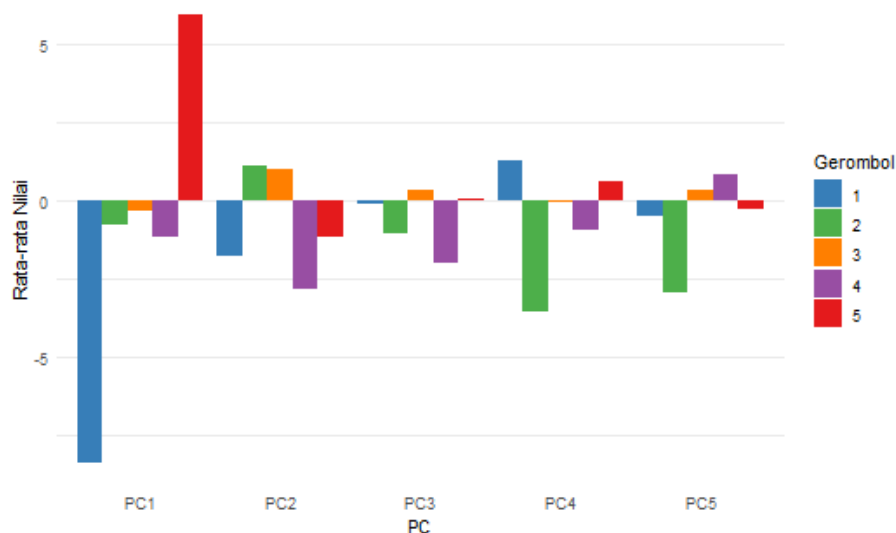
Keanggotaan tiap kelompok berdasarkan hasil pengelompokan dengan algoritma Rmixmod terdapat pada Tabel 7 dan visualisasi tiga dimensi pada Gambar 8.

**Tabel 7.** Jumlah anggota tiap kelompok

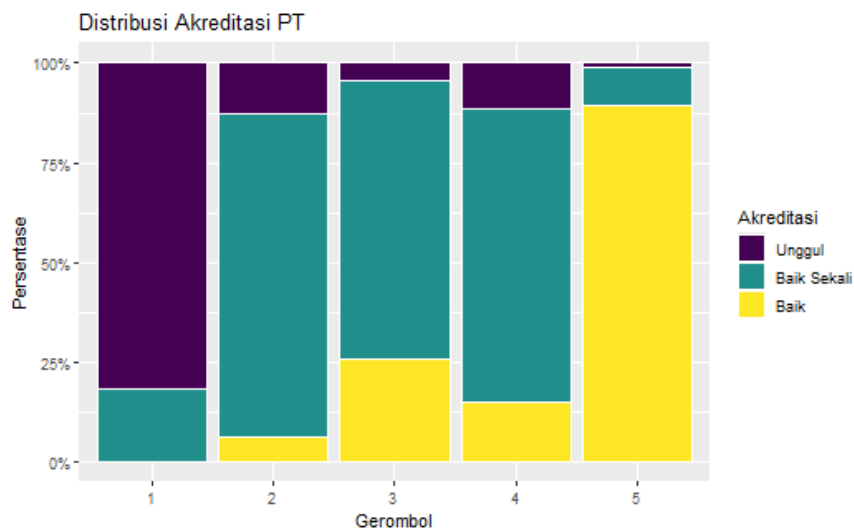
Kelompok	Anggota	Persentase (%)
1	87	10,83
2	32	3,99
3	468	58,28
4	53	6,60
5	163	20,30



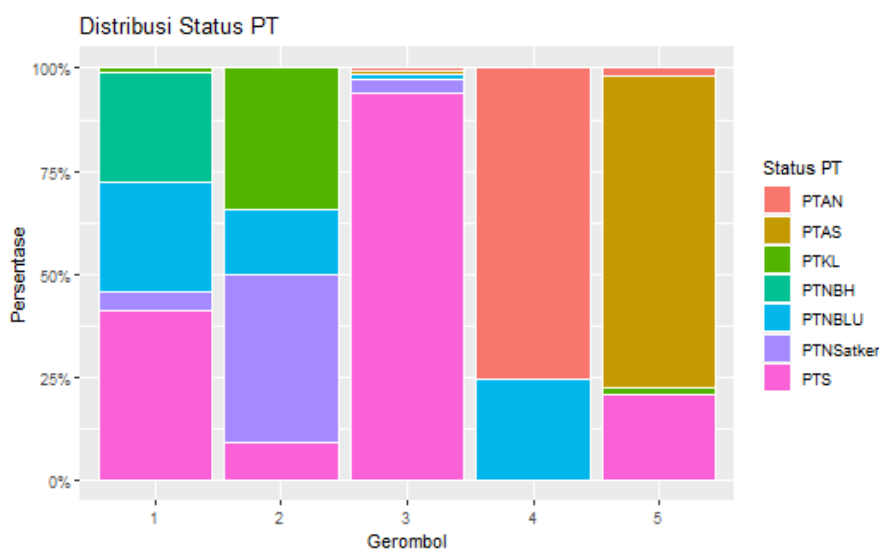
**Gambar 8.** Visualisasi Tiga Dimensi Hasil Pengelompokan



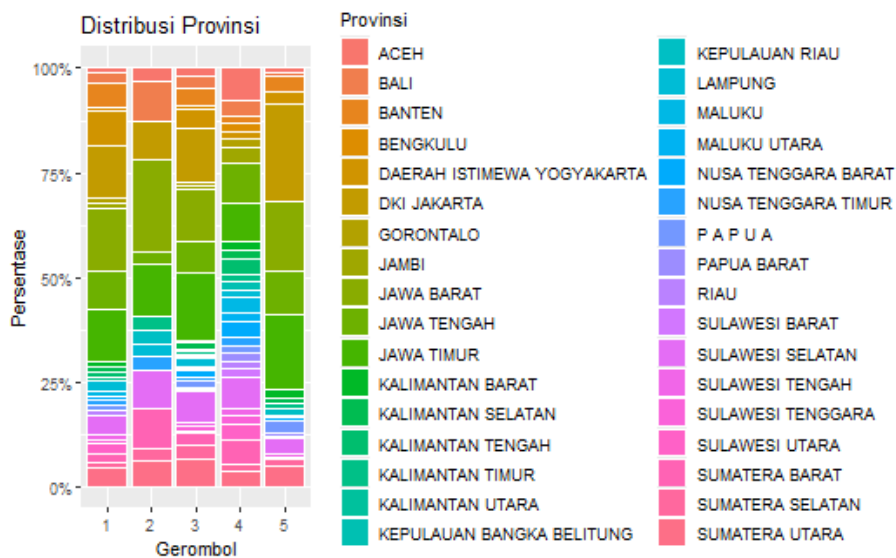
**Gambar 9.** Rata-rata Nilai PC pada Tiap Kelompok



**Gambar 10.** Distribusi Akreditasi Perguruan Tinggi per Kelompok



**Gambar 11.** Distribusi Status Perguruan Tinggi per Kelompok



**Gambar 12.** Distribusi Provinsi per Kelompok

Karakteristik kelompok berdasarkan Tabel 3, Gambar 9, Gambar 10, Gambar 11, dan Gambar 12 sebagai berikut:



- a. Kelompok 1 terdiri dari perguruan tinggi dengan status PTS (41%), PTNBH (26%), PTNBLU (26%) dengan akreditasi Unggul (82%) dan Baik Sekali (18%). Persebaran perguruan tinggi meliputi provinsi Jawa Barat (15%), DKI Jakarta (13%), Jawa Timur (13%). Perguruan tinggi dicirikan dengan total dosen sangat besar, *sinta score* yang sangat tinggi dan prodi pascasarjana yang beragam.
- b. Kelompok 2 terdiri dari perguruan tinggi berstatus PTKL (34%), PTNSatker (41%), PTNBLU (16%), dan PTS (9%). Akreditasi perguruan tinggi Baik Sekali (81%), Unggul (13%), dan Baik (6%). Persebaran perguruan tinggi meliputi provinsi Jawa Barat (22%), Jawa Timur (12%), DKI Jakarta (9%). Berbeda dengan kelompok lainnya, pada kelompok ini sebagian besar terdiri dari politeknik dengan keunggulan dalam *sinta score*, produktivitas 3Yr dan jumlah prodi S2 terapan.
- c. Kelompok 3 terdiri dari perguruan tinggi dengan status PTS (94%), akreditasi Baik Sekali (70%), Baik (26%), dan Unggul (4%). Persebaran provinsi yakni Jawa Timur (16%), DKI Jakarta (13%), Jawa Barat (13%). Perguruan tinggi memiliki *sinta score* dan produktivitas 3Yr yang relatif tinggi dan sebanding dengan jumlah *author*.
- d. Kelompok 4 terdiri dari perguruan tinggi dengan status PTAN (75%) dan PTNBLU (25%) dengan akreditasi perguruan tinggi Baik Sekali (74%), Baik (15%), dan Unggul (11%). Persebaran perguruan tinggi meliputi provinsi Jawa Tengah (9%), Jawa Timur (9%), Aceh (8%). Perguruan tinggi pada kelompok ini sebagian besar berada di bawah koordinasi Kementerian Agama (98%). Memiliki total dosen yang relatif besar namun tidak diimbangi dengan produktivitas penelitian, publikasi, dan pengabdian kepada masyarakat.
- e. Kelompok 5 terdiri dari perguruan tinggi berstatus PTAS (75%), PTS (21%) dengan akreditasi Baik (90%), Baik Sekali (9%), dan Unggul (1%). Persebaran perguruan tinggi meliputi provinsi DKI Jakarta (23%), Jawa Timur (18%), Jawa Barat (17%). Anggota kelompok ini berupa perguruan tinggi dengan kapasitas terbatas.

#### 4. KESIMPULAN

Penelitian ini membandingkan algoritma pengelompokan berbasis model yakni Rmixmod dan KAMILA dalam mengelompokkan perguruan tinggi di Indonesia berdasarkan kinerja perguruan tinggi. Data yang digunakan berupa data campuran yang terdiri dari data PCA hasil reduksi peubah numerik yang digabungkan dengan data kategorik. Penentuan jumlah kelompok optimal berdasarkan nilai BIC dan koefisien *Silhouette* sedangkan evaluasi algoritma pengelompokan berdasarkan nilai koefisien *Silhouette* maksimum, indeks Calinski-Harabasz maksimum, dan indeks Davies-Bouldin minimum. Hasil evaluasi pengelompokan diperoleh algoritma Rmixmod sebagai algoritma pengelompokan terbaik dengan nilai koefisien *Silhouette* sebesar 0,2878, indeks Calinski-Harabasz sebesar 253,9433, dan indeks Davies-Bouldin sebesar 1,5321 serta jumlah kelompok optimal sebanyak 5 kelompok. Kelompok 1 memiliki capaian sangat tinggi, kelompok 2 memiliki capaian tinggi, kelompok 2 memiliki capaian sedang, kelompok 4 memiliki capaian kurang optimal, dan kelompok 5 memiliki capaian terbatas. Hasil pengelompokan diharapkan dapat menjadi landasan dalam pembuatan kebijakan strategis yang efektif serta terdiferensiasi sesuai dengan karakteristik tiap kelompok perguruan tinggi.

#### REFERENCES

- [1] M. Mulyoto, U. Rosyidi, and R. Rugayah, "Mutu Perguruan Tinggi: Perspektif Peringkat Universitas Global dan Akreditasi Perguruan Tinggi di Indonesia," *Manajemen Pendidikan*, vol. 18, no. 1, pp. 26–41, Jul. 2023, doi: 10.23917/jmp.v18i1.20955.
- [2] M. F. Rouf, A. N. R. Attamimi, D. A. V. Putri, I. Nirmala, and A. N. Fadhilah, *Statistik Pendidikan Tinggi Tahun 2023*. Jakarta: Sekretariat Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, 2023.
- [3] J. Singh and D. Singh, "A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects," *Advanced Engineering Informatics*, vol. 62, Part C, no. 1, p. 102799, Oct. 2024, doi: 10.1016/J.AEI.2024.102799.
- [4] Q. Wen, "Application of Clustering Algorithm in Corporate Strategy and Risk," *Comput Intell Neurosci*, vol. 2022, no. 1, p. 8803375, Jan. 2022, doi: https://doi.org/10.1155/2022/8803375.
- [5] C. Zhang, S. Lasaulce, M. Hennebel, L. Saludjian, P. Panciatici, and H. V. Poor, "Decision-making oriented clustering: Application to pricing and power consumption scheduling," *Appl Energy*, vol. 297, no. 1, p. 117106, Sep. 2021, doi: 10.1016/j.apenergy.2021.117106.
- [6] E. Puspaputri *et al.*, *Panduan Penelitian dan Pengabdian kepada Masyarakat Tahun 2024*. Jakarta: Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, 2024.
- [7] M. Marhayati, A. M. Fa'ani, S. U. Ruhmanasari, and S. Faridah, "Application of K-Means Cluster Analysis for Grouping State Islamic University in Indonesia based on the Readiness Indicators for World Class University (WCU)," *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 8, no. 2, pp. 30–48, Nov. 2023, doi: 10.18860/ca.v8i2.18046.
- [8] V. Fatmawaty, I. Riadi, and H. Herman, "Higher Education Institution Clustering Based on Key Performance Indicators using Quartile Binning Method," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 1, Nov. 2024, doi: https://doi.org/10.30812/matrik.v24i1.4244.
- [9] A. D. Cipta, "Klasterisasi Perguruan Tinggi Swasta Berdasarkan Minat Siswa Menggunakan Metode K-Medoids: Andika Dwi Cipta, Asep Id Hadiana, Fajri Rahmat Umbara," *Journal of Informatics and Communication Technology (JICT)*, vol. 4, no. 2, pp. 21–29, Jan. 2023, doi: 10.52661/j\_ict.v4i2.116.
- [10] L. Alexandre, R. S. Costa, and R. Henriques, "DISA tool: Discriminative and informative subspace assessment with categorical and numerical outcomes," *PLoS One*, vol. 17, no. 10, p. e0276253, Oct. 2022, doi: 10.1371/journal.pone.0276253.



- [11] Y. Lee, C. Park, and S. Kang, “Deep Embedded Clustering Framework for Mixed Data,” *IEEE Access*, vol. 11, no. 1, pp. 33–40, 2023, doi: 10.1109/ACCESS.2022.3232372.
- [12] B. Ghattas and A. Sanchez San-Benito, “Clustering Approaches for Mixed-Type Data: A Comparative Study,” *J Probab Stat*, vol. 2025, no. 1, Jan. 2025, doi: 10.1155/jpas/2242100.
- [13] A. Ahmad and S. S. Khan, “Survey of State-of-the-Art Mixed Data Clustering Algorithms,” *IEEE Access*, vol. 7, no. 1, pp. 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.
- [14] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, Eds., “Model-based Clustering: Basic Ideas,” in *Model-Based Clustering and Classification for Data Science: With Applications in R*, in Cambridge Series in Statistical and Probabilistic Mathematics. , Cambridge: Cambridge University Press, 2019, pp. 15–78. doi: 10.1017/9781108644181.003.
- [15] P. Giordani, M. B. Ferraro, and F. Martella, “Model-Based Clustering,” in *An Introduction to Clustering with R*, P. Giordani, M. B. Ferraro, and F. Martella, Eds., Singapore: Springer Singapore, 2020, pp. 215–289. doi: 10.1007/978-981-13-0553-5\_6.
- [16] R. Lebrecht, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert, “Rmixmod: The R Package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library,” *J Stat Softw*, vol. 67, no. 6, 2015, doi: 10.18637/jss.v067.i06.
- [17] J. Roche *et al.*, “GenoTripto: A SNP genotype calling method for triploids,” *PLoS Comput Biol*, vol. 20, no. 9, p. e1012483, Sep. 2024, doi: 10.1371/journal.pcbi.1012483.
- [18] A. Foss, M. Markatou, B. Ray, and A. Heching, “A semiparametric method for clustering mixed data,” *Mach Learn*, vol. 105, no. 3, pp. 419–458, Dec. 2016, doi: 10.1007/s10994-016-5575-7.
- [19] J. Jimeno, M. Roy, and C. Tortora, “Clustering Mixed-Type Data: A Benchmark Study on KAMILA and K-Prototypes,” in *Data Analysis and Rationality in a Complex World*, T. Chadjipadelis, B. Lausen, A. Markos, T. R. Lee, A. Montanari, and R. Nugent, Eds., Cham: Springer International Publishing, 2021, pp. 83–91.
- [20] A. H. Foss and M. Markatou, “kamila: Clustering mixed-type data in R and hadoop,” *J Stat Softw*, vol. 83, no. 13, 2018, doi: 10.18637/jss.v083.i13.
- [21] “SINTA - Science and Technology Index.” Accessed: Jan. 11, 2025. [Online]. Available: <https://sinta.kemdikbud.go.id/>
- [22] “Badan Akreditasi Nasional Perguruan Tinggi – BAN-PT.” Accessed: Mar. 31, 2024. [Online]. Available: <https://www.banpt.or.id/>
- [23] “Website PPK BLU.” Accessed: Apr. 03, 2024. [Online]. Available: <https://blu-djpb.kemenkeu.go.id/>
- [24] “PDDikti.” Accessed: May 29, 2024. [Online]. Available: <https://pddikti.kemdiktisaintek.go.id/>
- [25] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Information Fusion*, vol. 59, no. 1, pp. 44–58, 2020, doi: <https://doi.org/10.1016/j.inffus.2020.01.005>.
- [26] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
- [27] S. Geng, “Analysis of the Different Statistical Metrics in Machine Learning,” in *Highlights in Science, Engineering and Technology IFMPT*, Darcy & Roy Press, 2024, pp. 350–356.
- [28] A. Aditya, B. N. Sari, and T. N. Padilah, “Comparison analysis of Euclidean and Gower distance measures on k-medoids cluster,” *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 1–7, Jan. 2021, doi: 10.14710/jtsiskom.2020.13747.
- [29] 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, Dec. 2022, doi: 10.1186/s40537-022-00639-7.