

OPTIMALISASI PEMETAAN KLASSTER KEPENDUDUKAN PROVINSI DI INDONESIA MENGGUNAKAN *ELBOW METHOD* DAN *DAVIES-BOULDIN INDEX*

Nilam Tri Astuti¹, Sunardi², Abdul Fadlil³

¹Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta

²Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Yogyakarta

³Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Yogyakarta

¹2408048024@webmail.uad.ac.id, ²sunardi@mti.uad.ac.id, ³fadlil@mti.uad.ac.id

Abstrak

Ketimpangan distribusi dan dinamika pertumbuhan penduduk antarprovinsi di Indonesia menimbulkan tantangan serius dalam perencanaan pembangunan yang berkeadilan dan berkelanjutan. Perbedaan jumlah penduduk, laju pertumbuhan, serta kontribusi terhadap total populasi nasional berimplikasi pada variasi kebutuhan infrastruktur, layanan publik, dan kapasitas wilayah dalam menopang aktivitas sosial-ekonomi. Oleh karena itu, diperlukan pendekatan analitik yang mampu mengelompokkan provinsi berdasarkan kesamaan karakteristik demografis guna mendukung perumusan kebijakan yang lebih terarah. Penelitian ini bertujuan memetakan pola kependudukan antarprovinsi di Indonesia menggunakan algoritma K-Means berdasarkan tiga variabel utama, yaitu rata-rata jumlah penduduk, laju pertumbuhan penduduk, dan persentase distribusi penduduk. Data sekunder yang digunakan bersumber dari Badan Pusat Statistik (BPS) periode 2016-2023 dan dirata-ratakan untuk merepresentasikan kondisi demografi jangka menengah yang relatif stabil. Seleksi variabel dilakukan melalui analisis korelasi, kemudian data dinormalisasi menggunakan Z-score untuk memastikan kesetaraan skala antarvariabel. Penentuan jumlah kluster optimal dilakukan dengan Elbow Method dan divalidasi menggunakan Davies-Bouldin Index (DBI). Hasil analisis menunjukkan titik siku pada $k = 4$ dan nilai DBI minimum sebesar 0,4430 pada $k = 4$, yang lebih rendah dibandingkan $k = 2$ (0,4838) dan $k = 3$ (0,8006), sehingga empat kluster dinilai sebagai struktur pengelompokan paling representatif. Klusterisasi menghasilkan empat kelompok demografis yang mencerminkan karakteristik berbeda, mulai dari provinsi dengan populasi sangat besar, wilayah pertumbuhan tinggi dengan jumlah penduduk relatif kecil, kelompok moderat, hingga wilayah dengan kepadatan ekstrem. Temuan ini menunjukkan bahwa pendekatan K-Means yang didukung validasi kuantitatif mampu menggambarkan heterogenitas demografi Indonesia secara lebih terstruktur serta berpotensi menjadi dasar dalam perencanaan pembangunan berbasis karakteristik wilayah.

Kata kunci: *Clustering, Algoritma K-Means, Analisis Demografi, Elbow Method, Davies-Bouldin Index(DBI)*

1. Pendahuluan

Pertumbuhan penduduk merupakan salah satu indikator fundamental dalam menentukan arah pembangunan sosial, ekonomi, dan tata ruang suatu negara. Peningkatan populasi, apabila tidak diiringi dengan pemerataan distribusi dan peningkatan kualitas sumber daya manusia, dapat menimbulkan ketimpangan sosial dan tekanan terhadap daya dukung lingkungan (Wirda et al., 2023). Distribusi penduduk yang tidak merata antarwilayah dapat menimbulkan perbedaan signifikan dalam kebutuhan layanan publik, infrastruktur, dan kapasitas ekologis suatu daerah (Nurparikha et al., 2023). Daerah dengan kepadatan penduduk tinggi umumnya menghadapi permasalahan keterbatasan ruang, kemacetan, dan degradasi lingkungan. Sedangkan daerah dengan kepadatan rendah sering tertinggal dalam akses pendidikan, kesehatan, dan kesempatan

ekonomi (Mu'awwanah & Ridho Illah, 2022). Oleh karena itu, pemerataan persebaran penduduk menjadi salah satu aspek penting dalam mewujudkan pembangunan yang inklusif dan berkelanjutan.

Dalam konteks Indonesia, variasi jumlah penduduk, kepadatan, dan laju pertumbuhan antarprovinsi menunjukkan ketimpangan tajam. Provinsi seperti DKI Jakarta, Jawa Barat, dan Jawa Timur menjadi wilayah dengan konsentrasi penduduk terbesar, sehingga menimbulkan tekanan tinggi terhadap pemanfaatan ruang dan infrastruktur perkotaan. Sebaliknya, wilayah seperti Kalimantan Utara, Papua Barat, dan Maluku Utara memiliki jumlah penduduk dan kepadatan yang rendah. Ketimpangan ini sejalan dengan fenomena kesenjangan pembangunan antarwilayah di Indonesia, dimana pertumbuhan ekonomi dan kepadatan aktivitas lebih terpusat di kawasan barat, sementara wilayah timur relatif tertinggal dalam

distribusi sumber daya dan kesempatan ekonomi (Sugiastuti & Pratama, 2022). Kondisi tersebut menunjukkan bahwa orientasi pembangunan yang terlalu berfokus pada pertumbuhan tanpa pemerataan berpotensi memperlebar disparitas antarwilayah dan memperlemah kohesi sosial ekonomi nasional.

Pemetaan karakteristik demografis secara akurat memiliki peran strategis bagi pemerintah dalam memastikan kebijakan pembangunan berjalan efektif dan tepat sasaran. Kajian empiris menunjukkan bahwa analisis kependudukan menjadi fondasi penting dalam proses perencanaan pembangunan karena mampu mengidentifikasi ketimpangan sosial dan menentukan prioritas intervensi di berbagai wilayah. Pendekatan ini tidak hanya berfungsi untuk pengalokasian sumber daya secara proporsional, tetapi juga mendukung perumusan kebijakan berbasis bukti yang adaptif terhadap dinamika sosial-ekonomi masyarakat. Oleh karena itu, penerapan metode analitik yang mampu mengelompokkan wilayah berdasarkan kesamaan karakteristik demografis. Serta menjadi sangat esensial dalam mendukung pengambilan keputusan pembangunan yang lebih objektif dan terarah (Asyafiyah et al., 2025).

Dalam beberapa tahun terakhir, berbagai penelitian mulai memanfaatkan analisis kluster untuk membaca pola-pola tersembunyi di balik data kependudukan dan pembangunan wilayah di Indonesia. Penelitian yang dilakukan oleh Novitasari et al. (2025) yaitu mengelompokkan provinsi berdasarkan Indeks Pembangunan Manusia menggunakan metode *K-Means* serta *Elbow* dan *Calinski-Harabasz Index*. Hasil penelitian menunjukkan bahwa klusterisasi mampu mengungkap perbedaan tingkat pembangunan antardaerah yang relevan bagi perumusan kebijakan publik. Siregar et al. (2024) mengkluster provinsi di Indonesia berdasarkan kepesertaan Jaminan Kesehatan Nasional dan menemukan variasi pola keikutsertaan antartahun, sehingga pemerintah dapat lebih peka terhadap wilayah yang membutuhkan penguatan intervensi jaminan sosial. Meskipun demikian, sebagian besar studi tersebut masih berfokus pada indikator tunggal atau terbatas, seperti IPM, kemiskinan, kualitas pemuda, atau indikator lingkungan tertentu, sehingga gambaran mengenai heterogenitas kependudukan antarprovinsi sebagai satu kesatuan nasional belum tergambar secara utuh. Di sisi lain, data kependudukan resmi yang diterbitkan Badan Pusat Statistik sesungguhnya menyimpan potensi besar untuk dibaca ulang secara lebih holistik melalui penggabungan beberapa indikator kunci, seperti jumlah penduduk, laju pertumbuhan, dan persentase kontribusi terhadap total populasi nasional. Celah ini membuka ruang bagi penelitian yang secara khusus menyoroti bagaimana provinsi-provinsi di Indonesia dapat dikelompokkan berdasarkan karakteristik demografis utamanya, sehingga pola ketimpangan dan kedekatan

karakter antarwilayah dapat dipahami dengan cara yang lebih sederhana namun tetap informatif bagi pengambil kebijakan.

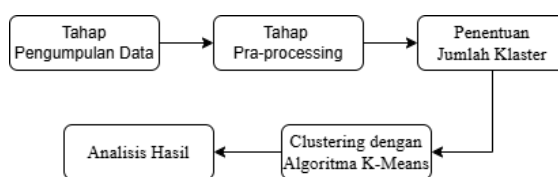
Klusterisasi (*clustering*) merupakan salah satu teknik analisis dalam *data mining* yang digunakan untuk mengelompokkan objek berdasarkan tingkat kemiripan fitur-fitur yang dimilikinya. Teknik ini berbeda dengan klasifikasi karena klusterisasi tidak membutuhkan label kelas sebelumnya dan sepenuhnya bergantung pada pola kesamaan antar data (Handayanna, 2023). Diantara berbagai metode klusterisasi, algoritma *K-Means* menjadi salah satu yang paling banyak digunakan karena kesederhanaannya, kecepatan proses, dan kemampuannya dalam mengolah data numerik multivariat untuk mengidentifikasi pola tersembunyi di dalam data (Aryanto et al., 2024). Metode ini telah banyak diterapkan dalam berbagai bidang, termasuk pengelompokan wilayah berdasarkan karakteristik sosial-ekonomi dan demografis yang dapat menghasilkan pembagian kelompok yang representatif sesuai dengan kedekatan nilai antar variabel.

Penelitian ini memanfaatkan data kependudukan dari Badan Pusat Statistik (BPS) dengan rentang tahun 2016-2023 yang meliputi jumlah penduduk, laju pertumbuhan penduduk, persentase distribusi penduduk, kepadatan penduduk, dan rasio jenis kelamin. Seluruh variabel tersebut kemudian dilakukan rata-rata untuk menangkap gambaran kondisi demografi jangka menengah. Penerapan metode *K-Means* dan evaluasi jumlah kluster optimal menggunakan *Elbow Method* dan *Davies-Bouldin Index* (DBI) dalam penelitian ini bertujuan untuk melakukan pemetaan kluster kependudukan provinsi di Indonesia sebagai dasar rekomendasi perencanaan pembangunan.

Hasil klusterisasi diharapkan dapat memberikan gambaran objektif mengenai karakteristik demografis antarprovinsi khususnya bagi Badan Perencanaan Pembangunan Nasional (Bappenas) sehingga dapat mendukung perumusan kebijakan pembangunan berbasis bukti (*evidence-based policy*). Temuan penelitian ini juga diharapkan mampu memberikan kontribusi terhadap literatur mengenai pemanfaatan *data mining* dalam kajian kependudukan dan pembangunan wilayah di Indonesia.

2. Metode

Penelitian ini melalui lima tahapan untuk melakukan *clustering* data seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Tahap Pengumpulan Data

Tahap pengumpulan data dalam penelitian ini dilakukan melalui pemanfaatan sumber data sekunder yang disediakan oleh BPS sebagai lembaga resmi pemerintah yang berwenang menyediakan data statistik nasional. Data yang digunakan berasal dari publikasi tabel statistik “Jumlah Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk, Kepadatan Penduduk, dan Rasio Jenis Kelamin Penduduk Menurut Provinsi” yang dapat diakses melalui laman resmi BPS www.bps.go.id. Dataset tersebut dipilih karena memuat indikator demografi yang lengkap, konsisten, dan tersedia dalam rentang waktu yang panjang sehingga memungkinkan analisis tren dan pengukuran rata-rata secara akurat. Pengumpulan data dilakukan untuk periode 2016-2023 dengan mempertimbangkan konsistensi penyajian variabel pada masing-masing tahun.

2.2 Tahap *Pre-processing*

Tahap *pre-processing* dilakukan untuk memastikan bahwa data berada dalam kondisi layak untuk dianalisis. Langkah-langkahnya meliputi pembersihan data, seleksi fitur, normalisasi data, penentuan jumlah kluster, klustering dengan algoritma K-Means, pengujian hasil klusterisasi dengan DBI dan analisis hasil.

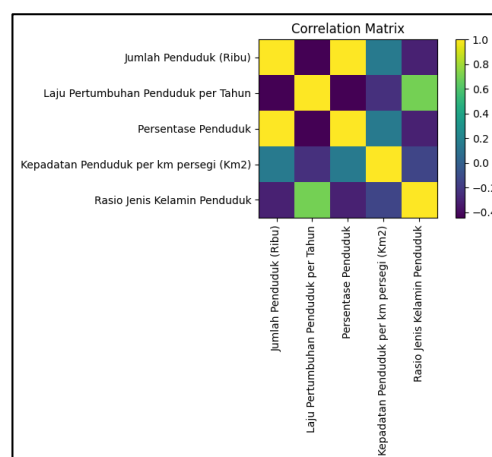
2.2.1 Pembersihan Data

Seluruh data diperiksa untuk memastikan tidak terdapat nilai kosong (*missing values*) atau duplikasi. Tahap ini sangat penting untuk menjaga integritas dataset, karena nilai hilang atau data ganda dapat menyebabkan distorsi dalam proses perhitungan jarak antarobjek yang menjadi dasar utama dalam algoritma *K-Means Clustering*. Proses pembersihan data (*data cleaning*) meliputi penghapusan data yang tidak relevan, pengubahan format data, serta normalisasi agar seluruh variabel memiliki skala yang seragam sebelum dilakukan proses klusterisasi (Kurniawan & Kurniawan, 2025). Data BPS bersifat tabel resmi sehingga tidak ditemukan inkonsistensi dalam format maupun nilai data.

2.2.2 Seleksi Fitur

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari BPS. Sumber data sekunder seperti BPS memiliki tingkat reliabilitas tinggi karena proses pengumpulan dan verifikasi dilakukan secara sistematis sesuai standar statistik nasional. Pada tahap awal, seluruh lima variabel demografis dianalisis untuk menentukan relevansinya terhadap proses

klusterisasi. Analisis ini bertujuan untuk memastikan bahwa setiap variabel memiliki kontribusi signifikan terhadap pembentukan kelompok yang representatif dalam konteks karakteristik kependudukan wilayah. Menurut Misbakhul Anam et al. (2025), pemilihan variabel yang tepat dalam penelitian berbasis *clustering* sangat menentukan akurasi hasil pengelompokan, karena variabel yang tidak relevan dapat menurunkan kualitas dan interpretabilitas kluster yang dihasilkan. Analisis tersebut dilakukan menggunakan *Heatmap Correlation* yang divisualisasikan seperti pada Gambar 2.



Gambar 2. Heatmap Correlation

Hasil *heatmap* menunjukkan bahwa jumlah penduduk dan persentase penduduk memiliki korelasi yang sangat kuat (mendekati +1), sehingga keduanya dianggap relevan untuk menggambarkan kontribusi populasi suatu provinsi terhadap populasi nasional. Sementara itu, variabel laju pertumbuhan penduduk meskipun tidak memiliki korelasi tinggi dengan variabel lain, tetap menjadi indikator penting untuk memahami dinamika kependudukan dan variasi perkembangan antarprovinsi. Sebaliknya, variabel kepadatan penduduk dan rasio jenis kelamin menunjukkan korelasi yang rendah terhadap variabel lain, sehingga pengaruhnya terhadap pembentukan kluster relatif kecil. Variabel tersebut berpotensi menghasilkan *noise* pada proses klasterisasi, sehingga tidak diikutsertakan dalam pemodelan. Dengan demikian, penelitian ini hanya menggunakan tiga variabel utama, yaitu:

1. Jumlah penduduk (jiwa)
2. Laju pertumbuhan penduduk (%)
3. Persentase penduduk (%)

Ketiga variabel tersebut menunjukkan relevansi statistika yang tinggi dan representatif terhadap profil demografis antarprovinsi. Selanjutnya, nilai rata-rata dari setiap variabel dihitung untuk masing-masing provinsi pada periode 2016-2023 agar menghasilkan gambaran jangka menengah yang lebih stabil. Pendekatan serupa juga diterapkan oleh Rachma et al. (2025), yang

melakukan proses perataan (*averaging*) variabel dalam studi klasterisasi untuk memperoleh representasi kondisi wilayah yang lebih objektif dan mengurangi fluktuasi tahunan data. Dengan demikian, penggunaan nilai rata-rata dinilai efektif untuk meningkatkan reliabilitas hasil klasterisasi berbasis data sekunder dari BPS.

2.2.3 Normalisasi Data

Setiap variabel memiliki skala yang berbeda sehingga diperlukan proses normalisasi agar tidak ada variabel yang mendominasi perhitungan jarak. Proses normalisasi pada penelitian ini menggunakan metode *Z-Score Normalization*, yang mentransformasi setiap nilai ke dalam satuan standar berdasarkan rata-rata dan standar deviasinya. Melalui metode ini, seluruh variabel akan memiliki rata-rata (*mean*) = 0 dan standar deviasi = 1. Pendekatan ini dipilih karena *Z-Score* efektif dalam mengatasi perbedaan skala pada variabel dengan rentang nilai yang berbeda besar. Dengan demikian, seluruh fitur yang digunakan dalam analisis berada pada skala yang sebanding sehingga proses klasterisasi dapat berlangsung secara optimal dan menghasilkan klaster yang lebih kompak serta terpisah dengan baik.

2.3 Penentuan Jumlah Klaster

Penentuan jumlah klaster merupakan tahapan krusial dalam proses klasterisasi karena pemilihan nilai k yang tidak tepat dapat menyebabkan pembentukan kelompok data yang tidak representatif, tumpang tindih, atau sulit diinterpretasikan secara analitis. Untuk memperoleh jumlah klaster yang paling sesuai, penelitian ini menerapkan dua pendekatan evaluatif, yakni *Elbow Method* dan DBI. Metode *Elbow* digunakan untuk mengidentifikasi titik optimal berdasarkan perubahan signifikan pada nilai *Within-Cluster Sum of Squares* (WCSS), sedangkan DBI berfungsi untuk menilai tingkat kedekatan dan keterpisahan antar-klaster melalui rasio dispersi intraklaster dan jarak antar *centroid*. Kombinasi kedua metode ini dinilai efektif dalam memastikan kualitas pengelompokan yang optimal, sebagaimana telah diterapkan dalam penelitian pengelompokan wilayah Indonesia berdasarkan Indeks Pembangunan Manusia menggunakan algoritma *K-Means* (Saputra & Muflih, 2025).

2.3.1 Elbow Method

Elbow Method merupakan salah satu teknik yang banyak digunakan dalam proses penentuan jumlah klaster optimal dengan memanfaatkan pengamatan terhadap nilai WCSS pada berbagai nilai k . Rumus dari elbow method dapat dilihat pada persamaan 1 (Amri et al., 2025):

$$WCSS = \sum_{k=1}^K \sum_{i \in C_k} |x_i - \mu_k|^2 \quad (1)$$

Keterangan:

k merupakan Jumlah klaster

C_k merupakan Klaster ke- k

x_i merupakan Titik data ke- i dalam klaster C_k

$|x_i - \mu_k|^2$ merupakan Jarak Euclidean kuadrat antara titik data x_i dan centeroid μ_k

Nilai WCSS merepresentasikan tingkat homogenitas data dalam suatu klaster. Semakin kecil nilai WCSS maka semakin baik kualitas pengelompokan yang dihasilkan. Proses analisis dilakukan dengan menghitung nilai WCSS secara bertahap untuk rentang $k=2$ hingga $k=9$, kemudian hasil perhitungannya divisualisasikan dalam bentuk grafik. Dari grafik tersebut, titik dimana penurunan nilai WCSS mulai melambat secara signifikan diidentifikasi sebagai titik siku (*elbow point*), yang menandakan bahwa penambahan jumlah klaster setelah titik tersebut tidak lagi memberikan peningkatan berarti terhadap kualitas hasil klasterisasi (Mayasari & Nugraha, 2023).

2.4 Clustering dengan Algoritma K-Means

Algoritma *K-Means* merupakan metode klasterisasi berbasis partisi yang digunakan untuk mengelompokkan data ke dalam sejumlah klaster yang telah ditentukan sebelumnya. Secara konseptual, algoritma ini bekerja dengan meminimalkan variasi di dalam klaster dan memaksimalkan perbedaan antar klaster melalui perhitungan jarak *Euclidean* (Iswahyudi & Rahmawati, 2025). Proses *K-Means* terdiri dari beberapa tahap inti. Pertama, jumlah klaster (k) ditentukan berdasarkan metode evaluasi tertentu, kemudian dilakukan inisialisasi centroid awal untuk meningkatkan stabilitas awal centroid. Kedua, setiap data dihitung jaraknya terhadap seluruh centroid, dan data tersebut dialokasikan ke klaster dengan jarak paling kecil. Ketiga, setelah seluruh data terkelompokkan, centroid diperbarui dengan menghitung rata-rata nilai dari semua anggota dalam klaster tersebut.

Tahapan alokasi dan pembaruan centroid berlangsung secara iteratif hingga mencapai kondisi konvergen. Konvergensi terjadi ketika perubahan posisi centroid antar iterasi sudah sangat kecil atau ketika tidak ada perpindahan anggota klaster pada iterasi berikutnya. Mekanisme iteratif ini memastikan bahwa algoritma secara bertahap menemukan struktur klaster yang paling representatif terhadap pola distribusi data. Pada akhirnya, *K-Means* mampu menghasilkan klaster yang kompak, terpisah dengan baik, dan stabil secara statistik. Efisiensi komputasi serta kemampuannya menangani data berukuran besar menjadikan algoritma ini banyak digunakan

dalam analisis demografi, termasuk pemetaan karakteristik populasi, identifikasi pola persebaran penduduk, dan klasifikasi wilayah berdasarkan indikator sosial-ekonomi yang kompleks (Febriansyah & Muntari, 2023).

2.5 Pengujian hasil klusterisasi dengan Davies-Bouldin Index (DBI)

Untuk memperkuat hasil analisis dari *Elbow Method*, penelitian ini juga menerapkan metrik *Davies-Bouldin Index* (DBI) sebagai metode validasi tambahan. DBI berfungsi untuk mengukur rasio antara dispersi intrakluster dan jarak antar-kluster, sehingga memberikan gambaran tentang seberapa baik suatu algoritma membentuk pemisahan antar kelompok data. Rumus dari DBI dapat dilihat pada persamaan 2 (Saputra & Muflih, 2025) :

$$DBI = \left(\frac{1}{N} \right) \sum_{i=1}^N \max_{j \neq i} (R_{ij}) \quad (2)$$

Keterangan:

N merupakan Total jumlah kluster

$\sum_{i=1}^N$ merupakan Hasil penjumlahan semua kluster dimulai dari kluster pertama ($i=1$) hingga kluster terakhir ($i=N$)

$\max_{j \neq i}$ merupakan Hasil perhitungan nilai maksimal dari rasio R_{ij} antar kluster I dan j.

R_{ij} merupakan rasio yang dipakai guna mengukur rasio kemiripan antara kluster I dan j, dinyatakan pada persamaan (3):

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (3)$$

Keterangan:

S_i dan S_j merupakan Jarak titik data pada kluster I atau j dengan centeroid

R_{ij} merupakan Jarak antara centeroid i dengan centeroid j

Nilai DBI dianggap optimal ketika berada pada titik minimum, karena kondisi tersebut menunjukkan bahwa kluster yang terbentuk memiliki tingkat kepadatan tinggi di dalam kelompok dan jarak yang signifikan antar kluster (Syaharani & Sriani, 2024).

2.6 Analisis Hasil

Analisis hasil klusterisasi dilakukan untuk memahami karakteristik setiap kluster yang terbentuk dan relevansinya terhadap kondisi demografis provinsi di Indonesia. Setelah proses *K-Means* menghasilkan kluster, tahap analisis dimulai dengan mengevaluasi distribusi provinsi pada masing-masing

kluster dan mengamati pola kedekatan berdasarkan tiga variabel utama, yaitu jumlah penduduk, laju pertumbuhan penduduk, dan persentase distribusi penduduk. Setiap kluster dibandingkan berdasarkan nilai *centroid* untuk mengidentifikasi ciri dominan yang membedakannya dari kluster lain, seperti kelompok provinsi dengan jumlah penduduk besar, wilayah dengan pertumbuhan tinggi, atau daerah yang memiliki kontribusi populasi kecil terhadap total nasional. Selain itu, digunakan visualisasi *scatter plot* untuk melihat penyebaran provinsi pada ruang multidimensi yang direduksi ke dalam dua variabel sehingga pola kluster dapat diamati secara visual.

Hasil kluster kemudian diinterpretasikan dalam konteks perencanaan pembangunan wilayah, sehingga setiap kluster dapat memberikan informasi terkait prioritas kebijakan, tekanan penduduk, dan arah perkembangan demografis. Pendekatan analisis ini memungkinkan hasil klusterisasi tidak hanya bersifat deskriptif, tetapi juga dapat memberikan pemahaman substantif mengenai struktur demografi antarprovinsi, sesuai tujuan penelitian dalam mendukung perumusan kebijakan berbasis data.

3. Hasil dan Pembahasan

3.1 Analisis Data

Sebelum proses normalisasi dilakukan, seluruh data demografis terlebih dahulu dihitung nilai rata-ratanya untuk setiap provinsi dalam rentang waktu 2016-2023. Tabel 1 menyajikan rata-rata untuk setiap variabel yang sudah diolah sebelum dilakukan proses klusterisasi.

Tabel 1. Rata-rata jumlah penduduk, laju pertumbuhan penduduk dan presentase penduduk

Provinsi	Rata-rata Jumlah Penduduk	Rata-rata Laju Pertumbuhan Penduduk	Rata-rata Presentase Penduduk
Aceh	5304,6875	1,72125	1,9725
Sumatera Utara	14697,5375	1,28625	5,46625
Sumatera Barat	5489,6	1,25625	2,0425
Riau	6636,3125	2,13125	2,46875
Jambi	3576,55	1,555	1,33
....			
....			
Maluku	1818,825	1,53875	0,675
Maluku Utara	1265,2625	1,92125	0,47125
Papua Barat	1045,925	2,63625	0,38875
Papua	3841,85	2,05125	1,425

Proses perataan (*averaging*) ini dilakukan untuk memperoleh representasi jangka menengah dari kondisi kependudukan tiap provinsi sehingga fluktuasi tahunan yang terlalu tinggi tidak memengaruhi stabilitas hasil klusterisasi. Pendekatan ini penting karena tren demografi bersifat dinamis, sehingga penggunaan nilai rata-rata mampu

memberikan gambaran yang lebih konsisten dan tidak bias oleh anomali tahunan.

Setelah nilai rata-rata dihitung, ketiga variabel utama berupa jumlah penduduk, laju pertumbuhan penduduk, dan persentase penduduk kemudian dinormalisasi menggunakan metode *Z-Score Normalization*. Normalisasi ini mengubah setiap variabel ke dalam satuan standar dengan rata-rata 0 dan standar deviasi 1. Tabel 2 menyajikan hasil normalisasi dengan *Z-score*.

Tabel 2. Hasil normalisasi dengan *Z-score*

Provinsi	Rata-rata Jumlah Penduduk	Rata-rata Laju Pertumbuhan Penduduk	Rata-rata Presentase Penduduk
Aceh	-0,235031	0,256043	-0,242911
Sumatera Utara	0,612890	-0,623859	-0,201648
Sumatera Barat	-0,218338	-0,684542	-0,228359
Riau	-0,114821	1,085376	-0,249013
Jambi	-0,391035	-0,080241	-0,250515
....			
....			
Maluku	-0,549710	-0,113111	-0,262814
Maluku Utara	-0,599682	0,660596	-0,262674
Papua Barat	-0,619482	2,106871	-0,273611
Papua	-0,367086	0,923555	-0,272954

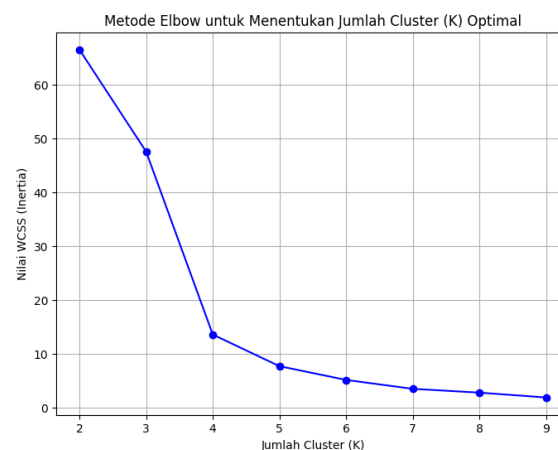
Hasil normalisasi menunjukkan bahwa sebagian besar provinsi berada pada rentang *Z-score* antara -0,6 hingga +0,6, mengindikasikan bahwa distribusi data telah seragam dan tidak ada variabel yang mendominasi skala perhitungan. Namun, beberapa provinsi dengan jumlah penduduk sangat besar menunjukkan nilai *Z-score* yang jauh di atas rata-rata, berada pada kisaran lebih dari +2. Hal ini mencerminkan adanya disparitas populasi yang signifikan antarwilayah. Sebaliknya, provinsi dengan populasi relatif kecil memiliki nilai *Z-score* negatif yang cukup rendah. Variasi inilah yang kemudian menjadi dasar penting dalam pembentukan pola klusterisasi. Untuk laju pertumbuhan penduduk, sebagian besar provinsi berada dalam kisaran *Z-score* -1 hingga +1, menunjukkan bahwa pertumbuhan penduduk di sebagian besar wilayah relatif homogen. Provinsi dengan pertumbuhan sangat tinggi seperti Kepulauan Riau, Kalimantan Utara, dan Papua Barat memiliki *Z-score* positif yang menonjol.

Proses normalisasi ini memastikan bahwa seluruh variabel berkontribusi secara proporsional dalam perhitungan jarak *Euclidean* pada algoritma *K-Means*. Dengan demikian, struktur klaster yang dihasilkan menjadi lebih akurat, kompak, dan dapat merepresentasikan kondisi demografi antarprovinsi secara lebih objektif.

3.2 Penentuan Jumlah Klaster

Penentuan jumlah klaster optimal dalam penelitian ini dilakukan melalui dua pendekatan, yaitu *Elbow Method* dan DBI. Kedua metode ini digunakan secara bersamaan untuk memastikan bahwa jumlah klaster yang dipilih tidak hanya didasarkan pada pengamatan visual, tetapi juga

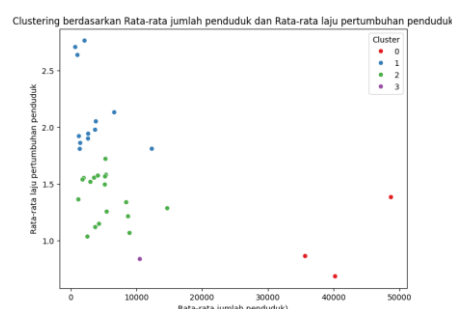
didukung oleh validasi kuantitatif terhadap kualitas pemisahan klaster. *Elbow Method* diterapkan dengan menghitung nilai *WCSS* untuk setiap nilai *k* pada rentang 2 hingga 9. Hasil perhitungan tersebut kemudian divisualisasikan dalam grafik seperti pada Gambar 3.

Gambar 3. Grafik *Elbow Method*

Grafik menunjukkan penurunan *WCSS* yang sangat drastis dari $k = 2$ ke $k = 3$, dan kembali menurun tajam pada $k = 4$. Setelah $k = 4$, penurunan *WCSS* menjadi semakin landai dan tidak menunjukkan penurunan signifikan. Pola tersebut mengindikasikan bahwa titik siku (*elbow*) berada pada $k = 4$, karena pada titik inilah penambahan jumlah klaster selanjutnya tidak memberikan peningkatan kualitas pemisahan klaster yang berarti. Dengan demikian, *Elbow Method* merekomendasikan jumlah klaster optimal sebanyak empat.

3.3 Clustering dengan algoritma *K-means*

Penerapan algoritma *K-Means* dengan jumlah klaster optimal sebanyak empat klaster menghasilkan pemetaan demografis yang memperlihatkan perbedaan karakteristik antarprovinsi di Indonesia secara lebih terstruktur. Pengelompokan dilakukan berdasarkan tiga variabel utama, yaitu rata-rata jumlah penduduk, rata-rata laju pertumbuhan penduduk, dan persentase penduduk. Untuk mempermudah visualisasi, hasil klusterisasi ditampilkan pada *scatter plot* dua dimensi dengan menggunakan dua variabel, yaitu rata-rata jumlah penduduk dan laju pertumbuhan penduduk, seperti terlihat pada Gambar 4.



Gambar 4. Hasil Klusterisasi

Berdasarkan hasil visualisasi di atas menunjukkan bahwa klaster yang terbentuk memiliki pola penyebaran yang jelas dan tidak saling tumpang tindih secara ekstrem. Hal ini mengindikasikan bahwa ketiga variabel yang digunakan mampu memisahkan karakteristik demografis antarprovinsi. Setiap warna pada grafik merepresentasikan nomor klaster yang dihasilkan model. Dari grafik tersebut terlihat bahwa provinsi dengan jumlah penduduk besar cenderung terkelompok pada klaster tertentu, sementara provinsi berpenduduk kecil tersebar pada klaster dengan pola pertumbuhan yang berbeda. Klasterisasi terhadap wilayah-wilayah dapat dilihat pada Tabel 3.

Tabel 3. Daftar wilayah per klaster

Klaster	Provinsi
0	Jawa Barat, Jawa Tengah, Jawa Timur
1	Riau, Kepulauan Bangka Belitung, Kepulauan Riau, Banten, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Sulawesi Tenggara, Sulawesi Barat, Maluku Utara, Papua Barat, Papua
2	Aceh, Sumatera Utara, Sumatera Barat, Jambi, Sumatera Selatan, Bengkulu, Lampung, DI Yogyakarta, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Selatan, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Gorontalo, Maluku
3	DKI Jakarta

Daftar provinsi dikelompokkan kedalam empat klaster berdasarkan hasil analisis *K-Means*. Klaster 0 terdiri dari 3 provinsi, klaster 1 mencakup 12 provinsi, klaster 2 beranggotakan 18 provinsi, dan klaster 3 terdiri dari 1 provinsi. Secara keseluruhan, seluruh provinsi di Indonesia berhasil dipetakan kedalam empat kelompok dengan karakteristik demografis yang berbeda.

3.4 Pengujian hasil klasterisasi dengan *Davies-Bouldin Index (DBI)*

Berdasarkan hasil klasterisasi kemudian dilakukan validasi menggunakan DBI yang mengukur tingkat kemiripan intraklaster dan keterpisahan antar klaster. Nilai DBI yang lebih kecil menunjukkan kualitas klaster yang lebih baik. Hasil perhitungan DBI dapat dilihat pada Tabel 4.

Tabel 4. Hasil Perhitungan DBI

Nilai k	Hasil
2	0.4838
3	0.8006
4	0.4993
5	0.4430
6	0.4645
7	0.3965
8	0.4448
9	0.3896

Secara umum, hasil DBI menunjukkan bahwa semakin besar nilai k maka semakin kecil nilai indeks yang dihasilkan. Namun demikian, penentuan jumlah klaster tidak hanya mempertimbangkan nilai DBI terkecil, tetapi juga harus menjaga keseimbangan antara kualitas klaster dan interpretabilitas hasil.

Dalam konteks penelitian demografi, jumlah klaster yang terlalu besar dapat menghasilkan pengelompokan yang kurang bermakna dan sulit diinterpretasikan. Meskipun DBI terkecil berada pada $k = 9$, nilai-nilai pada $k = 4$ hingga $k = 7$ menunjukkan perbedaan yang relatif kecil (rentang 0,39–0,49). Oleh karena itu, nilai $k = 4$ dipilih sebagai jumlah klaster optimal karena menawarkan kompromi terbaik antara kualitas klaster yang cukup baik (DBI = 0,4993), stabilitas struktur klaster, dan interpretabilitas hasil dalam konteks analisis kependudukan Indonesia. Dengan mempertimbangkan kedua metode tersebut, penelitian ini menetapkan bahwa model klasterisasi yang paling optimal menggunakan empat klaster.

3.5 Analisis Hasil

Hasil penerapan algoritma *K-Means* pada data kependudukan provinsi di Indonesia tahun 2016–2023 menunjukkan bahwa empat klaster merupakan struktur pengelompokan yang paling optimal, sebagaimana ditunjukkan oleh titik siku pada *Elbow Method* dan nilai minimum *Davies-Bouldin Index* sebesar 0,4430 pada $k = 4$. Nilai DBI yang relatif rendah ini mengindikasikan bahwa klaster yang terbentuk memiliki tingkat kekompakan internal yang baik dan jarak antarklaster yang cukup tegas, sehingga pola perbedaan karakteristik demografis antarprovinsi dapat dibaca dengan lebih jelas. Temuan ini sejalan dengan penelitian (Fahmiah & Ningrum, 2023) yang juga memperoleh struktur klaster yang stabil ketika menggunakan *K-Means* untuk mengelompokkan provinsi di Indonesia berdasarkan Indeks Pembangunan Manusia, di mana pemilihan jumlah klaster yang tepat terbukti penting untuk menghasilkan pemetaan wilayah yang informatif bagi perumusan kebijakan.

Klaster 0 terdiri dari provinsi dengan jumlah penduduk sangat besar, seperti Jawa Barat, Jawa Tengah, dan Jawa Timur. Klaster ini ditandai oleh nilai rata-rata jumlah penduduk yang jauh melampaui provinsi lainnya, bahkan mencapai puluhan juta jiwa. Meskipun demikian, laju pertumbuhan penduduk pada klaster ini cenderung rendah hingga moderat. Kondisi ini mengindikasikan bahwa wilayah berpenduduk besar umumnya telah mencapai fase pertumbuhan yang lebih stabil. Struktur kependudukan seperti ini sering berkaitan dengan tingkat urbanisasi tinggi, tekanan penggunaan lahan, dan kebutuhan pelayanan publik yang besar. Keberadaan provinsi dalam klaster ini menunjukkan pusat-pusat konsentrasi demografis yang menjadi penggerak aktivitas ekonomi nasional.

Klaster 1 merupakan kelompok provinsi yang memiliki jumlah penduduk rendah tetapi laju pertumbuhan penduduk tinggi. Provinsi seperti Papua, Papua Barat, Kalimantan Utara, dan Kepulauan Riau termasuk dalam klaster ini. Karakteristik utama klaster ini adalah dinamika pertumbuhan penduduk yang sangat cepat, meskipun kontribusinya terhadap total populasi nasional masih kecil. Pertumbuhan yang tinggi ini dapat dipengaruhi oleh migrasi masuk, pengembangan wilayah baru, atau peningkatan sosial ekonomi di kawasan timur dan perbatasan Indonesia. Klaster ini menandakan wilayah yang berpotensi mengalami transformasi demografis signifikan pada masa mendatang dan memerlukan perencanaan jangka panjang terkait infrastruktur, pendidikan, dan pengembangan wilayah.

Klaster 2 merupakan kelompok dengan jumlah anggota terbanyak dan mencerminkan provinsi dengan jumlah penduduk sedang serta laju pertumbuhan moderat. Provinsi-provinsi pada klaster ini tersebar di hampir seluruh wilayah Indonesia, termasuk Sumatera, Kalimantan, Sulawesi, Bali, dan Nusa Tenggara. Pola kependudukan yang relatif seimbang pada klaster ini menunjukkan bahwa wilayah tersebut memiliki tekanan penduduk yang tidak terlalu besar namun tetap berkembang secara stabil dari tahun ke tahun. Klaster ini dapat dianggap sebagai kelompok wilayah dengan kondisi demografis normal, dimana beban penduduk dan pertumbuhannya berada pada tingkat yang dapat dikendalikan. Pengembangan wilayah pada klaster ini umumnya diarahkan untuk pemerataan layanan publik, peningkatan produktivitas ekonomi daerah, dan peningkatan kualitas hidup masyarakat.

Klaster 3 hanya terdiri dari satu provinsi, yaitu DKI Jakarta, yang memiliki jumlah penduduk besar dan kepadatan penduduk ekstrem. DKI Jakarta memiliki nilai kepadatan penduduk yang jauh di atas provinsi lain karena dominasi kawasan perkotaan dan pusat pemerintahan. Laju pertumbuhan penduduk yang rendah menunjukkan bahwa penambahan penduduk di wilayah ini terjadi lebih banyak karena mobilitas penduduk harian daripada pertumbuhan alami. Kondisi demografis yang ekstrem ini membuat DKI Jakarta menjadi klaster tersendiri, mencerminkan wilayah metropolitan yang menghadapi tantangan kompleks seperti kemacetan, keterbatasan lahan, dan kebutuhan infrastruktur yang sangat tinggi.

Dari sisi metodologis, hasil penelitian ini memperkuat bukti bahwa kombinasi *K-Means* dengan indeks validasi seperti DBI efektif digunakan untuk menganalisis data kependudukan berskala besar dan multivariat. Pendekatan serupa digunakan dalam penelitian terbaru mengenai pengelompokan tingkat pengangguran provinsi di Indonesia dengan data BPS 2013-2023, di mana *K-Means* berhasil memisahkan provinsi ke dalam beberapa klaster berdasarkan kemiripan tingkat pengangguran

sehingga memudahkan penentuan prioritas kebijakan pasar kerja di masing-masing kelompok wilayah. Konsistensi keberhasilan *K-Means* pada berbagai konteks mulai dari IPM, kemiskinan, indikator kesehatan, hingga kependudukan mengindikasikan bahwa pemanfaatan metode ini pada data demografi nasional, seperti yang dilakukan dalam penelitian ini, memiliki landasan empiris yang kuat dan relevan untuk terus dikembangkan.

Secara keseluruhan, hasil klasterisasi menunjukkan bahwa struktur demografi Indonesia tidak seragam dan membentuk kelompok-kelompok yang mencerminkan tingkat perkembangan penduduk yang berbeda. Provinsi dengan populasi besar membentuk klaster tersendiri, begitu pula provinsi dengan pertumbuhan sangat tinggi atau yang memiliki kepadatan ekstrem. Sementara itu, sebagian besar provinsi berada dalam kategori menengah yang relatif stabil. Analisis ini memberikan gambaran menyeluruh mengenai karakteristik demografis Indonesia dan dapat menjadi acuan penting dalam penetapan prioritas pembangunan berdasarkan kebutuhan penduduk di setiap klaster.

4. Kesimpulan

Penelitian ini menunjukkan bahwa penerapan algoritma *K-Means* yang dikombinasikan dengan *Elbow Method* dan *Davies-Bouldin Index* mampu memetakan klaster kependudukan antarprovinsi di Indonesia secara lebih terstruktur, sehingga perbedaan karakter demografis seperti besaran populasi, laju pertumbuhan, dan distribusi penduduk dapat terlihat dengan jelas dan mudah dipahami oleh para pengambil kebijakan maupun masyarakat umum. Empat klaster yang terbentuk menggambarkan adanya kelompok provinsi dengan penduduk sangat besar, provinsi berpenduduk kecil hingga sedang dengan pertumbuhan tinggi, provinsi berpenduduk dan pertumbuhan moderat, serta kelompok khusus dengan kepadatan sangat tinggi seperti DKI Jakarta, yang secara tidak langsung menegaskan bahwa setiap wilayah memerlukan pendekatan pembangunan yang berbeda dan lebih peka terhadap konteks lokal. Temuan ini diharapkan dapat menjadi pijakan awal bagi perumusan kebijakan pembangunan yang lebih adil dan berkelanjutan, misalnya dalam penentuan prioritas infrastruktur, layanan dasar, serta strategi pengendalian kepadatan penduduk, sehingga manfaat pembangunan dapat dirasakan lebih merata oleh warga di berbagai provinsi. Penelitian ini masih dapat dikembangkan dengan menambahkan variabel sosial-ekonomi lain seperti tingkat kemiskinan, pendidikan, dan ketenagakerjaan, membandingkan beberapa algoritma clustering yang berbeda, ataupun memperluas skala analisis hingga level kabupaten/kota, sehingga peta klaster yang dihasilkan menjadi semakin kaya, mendalam, dan relevan sebagai bahan refleksi bersama dalam menyusun arah

pembangunan Indonesia yang lebih inklusif dan berorientasi pada kesejahteraan manusia.

Daftar Pustaka:

- Amri, Muh. A., Risal, A. A. N., Bakri, Muh. F., & Surianto, D. F. (2025). Mini-Batch K-Means Clustering Untuk Pengelompokan Kemandirian Daerah Di Sulawesi Selatan. *Jurnal Informatika Polinema*, 11(2), 235–244. <https://doi.org/10.33795/jip.v11i2.6871>
- Aryanto, R. P., Nilogiri, A., & Wardoyo, A. E. (2024). Klasterisasi Jumlah Penduduk Provinsi Jawa Timur Tahun 2021-2023 Menggunakan Algoritma K-Means. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 9(2), 134–146. <https://doi.org/10.14421/jiska.2024.9.2.134-146>
- Asyafiiyah, G. R. U., Widyastuti, A., & Andriani, F. (2025). Analisis Wilayah Prioritas Pembangunan di Provinsi Jawa Timur Berdasarkan Indikator Sosial Menggunakan Metode K-Means Clustering. *Arcitech: Journal of Computer Science and Artificial Intelligence*, 5(1), 32–48. <https://doi.org/10.29240/arcitech.v5i1.13488>
- Fahmiyah, I., & Ningrum, R. A. (2023). Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories. *Journal of Advanced Technology and Multidiscipline*, 2(1), 27–33. <https://doi.org/10.20473/jatm.v2i1.45070>
- Febriansyah, F., & Muntari, S. (2023). Penerapan Algoritma K-Means untuk Klasterisasi Penduduk Miskin pada Kota Pagar Alam. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 8(1), 66–77. <https://doi.org/10.14421/jiska.2023.8.1.66-77>
- Handayanna, F. (2023). Penerapan Algoritma K-Means Untuk Klasterisasi Penduduk Miskin Di Provinsi Banten. *Inti Nusa Mandiri*, 18(1), 93–99. <https://doi.org/10.33480/inti.v18i1.4399>
- Iswahyudi, L. E., & Rahmawati, F. (2025). Segmentasi Wilayah melalui K-Means Cluster untuk Mencapai SDGs 5 “Kesetaraan Gender.” *Jurnal Ilmu Sosial Dan Humaniora*, 14(2), 258–268. <https://doi.org/10.23887/jish.v14i2.92647>
- Kurniawan, W., & Kurniawan, R. (2025). Penerapan Algoritma K-Means Clustering Dalam Menentukan Peluang Masuk Siswa Ke Universitas Negeri. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, 7(1), 386–393. <https://doi.org/10.51401/jinteks.v7i1.5586>
- Mayasari, S. N., & Nugraha, J. (2023). Implementasi K-Means Cluster Analysis untuk Mengelompokkan Kabupaten/Kota Berdasarkan Data Kemiskinan di Provinsi Jawa Tengah Tahun 2022. *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, 3(2), 317–329. <https://doi.org/10.24002/konstelasi.v3i2.7200>
- Misbakhul Anam, Annisa Maulana Majid, & Ermanto. (2025). Analisis Pola Pengangguran Menggunakan Metode Clustering Algoritma K-Means Di Wilayah Kabupaten Cirebon. *Jurnal Teknologi Dan Manajemen Industri Terapan*, 4(4), 1691–1701. <https://doi.org/10.55826/jtmit.v4i4.743>
- Mu’awwanah, U., & Ridho Illah, G. R. (2022). Problematika Kependudukan Indonesia. *Al-Tsaman : Jurnal Ekonomi Dan Keuangan Islam*, 4(1), 63–78. <https://doi.org/10.62097/al-tsaman.v4i1.877>
- Novitasari, D., Harmaimun Hidayah, A., Kamila, R., & Nasrudin, M. (2025). *Journal of Artificial Intelligence and Engineering Applications Cluster Modeling with K-Means on Provincial Data in Indonesia Based on Environmental Indicators* (Vol. 4, Issue 3). <https://ioinformatic.org/>
- Nurparikha, D. A., Ramdani, E. M., Sudrajat, A. S., & Novira, A. (2023). PEMERATAAN Pembangunan Melalui Kebijakan Pemekaran Wilayah. *Jurnal Kebijakan Publik*, 14(4), 407. <https://doi.org/10.31258/jkp.v14i4.8340>
- Rachma, M., Budy Santoso, C., Informasi, S., Teknologi dan Desain, F., Pembangunan Jaya, U., & Selatan, T. (2025). Klasterisasi Data Pengangguran Di Pulau Jawa Menggunakan Algoritma K-Means Dalam Penanggulangan Pengangguran Tahun 2020-2023. In *Idealis: Indonesia Journal Information System* (Vol. 8, Issue 2). <https://doi.org/https://doi.org/10.36080/idealis.v8i2.3522>
- Saputra, N. R., & Muflih, G. Z. (2025). Pengelompokan Wilayah Indonesia Berdasarkan Komponen Indeks Pembangunan Manusia Dengan Pendekatan Algoritma K-Means Clustering. *Skanika: Sistem Komputer Dan Teknik Informatika*, 8(1), 156–167. <https://doi.org/10.36080/skanika.v8i1.3318>
- Siregar, D., Rahayu, W., Wardana, B. M., Kettrin Natasya Stefany, & Bayu Wibisono. (2024). Characteristics of Provinces in Indonesia Based on JKN Indicator Outcomes by Gaussian Mixture Model with Expectation-Maximization Algorithm and Biplot. *Jurnal Statistika Dan Aplikasinya*, 8(1), 17–30. <https://doi.org/10.21009/jsa.08102>
- Sugiasuti, R. H., & Pratama, M. R. (2022). *Dampak Buruk Pembangunan Tanpa Pemerataan: Kesenjangan Ekonomi Antar Wilayah Di Indonesia*. <https://doi.org/https://doi.org/10.21776/ub.prof.it.2022.-16.01.8>
- Syahrani, W., & Sriani, S. (2024). Application of k-Means Algorithm on Clustering Poor Population Data for Extreme Poverty

Elimination. *SISTEMASI*, 13(4), 1732.
<https://doi.org/10.32520/stmsi.v13i4.4384>
Wirda, F., Putra, T. J., Wardani, R., & Herizon, H.
(2023). Apakah Pertumbuhan Penduduk

Berdampak Bagi Pertumbuhan Ekonomi
Kabupaten Solok. *Amal: Jurnal Ekonomi
Syariah*, 5(01).
<https://doi.org/10.33477/eksy.v5i01.5476>