

CLUSTERING PROVINSI DI INDONESIA BERDASARKAN KARAKTERISTIK SOSIAL EKONOMI MENGGUNAKAN K-MEANS DAN DBSCAN

MUHAMMAD RIZQA SALAS



**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “*Clustering Provinsi di Indonesia Berdasarkan Karakteristik Sosial Ekonomi Menggunakan K-Means dan DBSCAN*” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dalam penyusunan karya ini, penulis menggunakan bantuan kecerdasan buatan Claude untuk membantu penyuntingan tata bahasa dan kejelasan kalimat, serta penyusunan sebagian kode analisis dan visualisasi data. Setelah menggunakan layanan tersebut, penulis meninjau dan menyunting konten sesuai kebutuhan serta bertanggung jawab penuh atas isi karya tugas akhir ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2026

Muhammad Rizqa Salas
G1401221060



ABSTRAK

MUHAMMAD RIZQA SALAS. *Clustering* Provinsi di Indonesia Berdasarkan Karakteristik Sosial Ekonomi Menggunakan K-Means dan DBSCAN. Dibimbing oleh SEPTIAN RAHARDIANTORO dan RAHMA ANISA.

Indonesia memiliki heterogenitas sosial ekonomi dan kesenjangan pembangunan antarwilayah, sehingga diperlukan segmentasi wilayah sebagai dasar kebijakan yang tepat sasaran. Permasalahan ini didekati melalui algoritma K-Means dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) yang diterapkan pada 17 indikator sosial ekonomi 38 provinsi dari Badan Pusat Statistik mencakup sektor ekonomi, pendidikan, kesehatan, dan teknologi; karena jumlah peubah yang banyak, data terlebih dahulu direduksi dimensinya melalui Analisis Komponen Utama per sektor, Analisis Faktor, dan *Uniform Manifold Approximation and Projection* (UMAP), dengan seleksi VIF dan data tanpa penanganan sebagai pembanding. Penelitian ini bertujuan mengelompokkan provinsi berdasarkan indikator tersebut serta membandingkan kinerja kedua algoritma dalam menghasilkan gerombol yang optimal. Kesepuluh kombinasi metode dievaluasi menggunakan *Silhouette Index*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). K-Means dengan *input* UMAP terpilih sebagai metode terbaik (*Silhouette Index* = 0,4554; DBI = 0,7604; CHI = 46,03) dan menghasilkan tiga gerombol dengan karakteristik berbeda: satu gerombol bernilai rendah pada mayoritas peubah sektor pembangunan yang menghimpun Kawasan Timur Indonesia dan wilayah perbatasan, satu gerombol bernilai menengah, serta satu gerombol bernilai tinggi yang menghimpun sebagian besar provinsi di Sumatra, Jawa, Kalimantan, dan Sulawesi Utara. Secara rata-rata, K-Means (*Silhouette Index* = 0,3344) cenderung lebih baik dibanding DBSCAN (0,3237) karena sensitivitasnya terhadap hyperparameter ϵ dan MinPts membuat beberapa kombinasi metode dengan DBSCAN pada kasus ini kurang mampu memisahkan karakteristik provinsi, sehingga hanya menghasilkan satu gerombol dan sisanya dianggap sebagai *noise*.

Kata kunci: DBSCAN, K-Means, penggerombolan, sosial ekonomi



ABSTRACT

MUHAMMAD RIZQA SALAS. Clustering of Indonesian Provinces Based on Socio-Economic Characteristics Using K-Means and DBSCAN. Supervised by SEPTIAN RAHARDIANTORO and RAHMA ANISA.

Indonesia has socio-economic heterogeneity and development gaps between regions, so regional segmentation is needed as a basis for targeted policies. This problem is approached through the K-Means algorithm and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which is applied to 17 socio-economic indicators in 38 provinces from the Central Statistics Agency covering the economic, education, health, and technology sectors; Due to the large number of variables, the data is first dimensionally reduced through Key Component Analysis per sector, Factor Analysis, and Uniform Manifold Approximation and Projection (UMAP), with VIF selection and unhandled data as a comparison. This study aims to group provinces based on these indicators and compare the performance of the two algorithms in producing optimal clusters. The ten combinations of methods were evaluated using the Silhouette Index, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index (CHI). K-Means with UMAP input selected as the best method (Silhouette Index = 0.4554; DBI = 0.7604; CHI = 46.03) and produced three gangs with different characteristics: one low-value gang in the majority of development sector variables that gathers the Eastern Region of Indonesia and border areas, one mid-value gang, and one high-value gang that gathers most of the provinces in Sumatra, Java, Kalimantan, and North Sulawesi. On average, K-Means (Silhouette Index = 0.3344) tends to be better than DBSCAN (0.3237) because of its sensitivity to ϵ hyperparameters and MinPts makes some combination methods with DBSCAN in this case less able to separate the provincial characteristics, resulting in only one cluster and the rest being considered as noise.

Keywords: clustering, DBSCAN, K-Means, socio-economic



@Hak cipta milik IPB University

IPB University



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026¹
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

CLUSTERING PROVINSI DI INDONESIA BERDASARKAN KARAKTERISTIK SOSIAL EKONOMI MENGGUNAKAN K-MEANS DAN DBSCAN

MUHAMMAD RIZQA SALAS

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Penguji pada ujian Skripsi:

1. Prof. Dr. Ir. Muhammad Nur Aidi, M.S.



Judul Skripsi : *Clustering Provinsi di Indonesia Berdasarkan Karakteristik Sosial
Ekonomi Menggunakan K-Means dan DBSCAN*

Nama : Muhammad Rizqa Salas

NIM : G1401221060

Disetujui oleh

Pembimbing 1:

Dr. Septian Rahardianto, S.Stat., M.Si.

Pembimbing 2:

Rahma Anisa, S.Stat., M.Act.Sc.

Diketahui oleh

Ketua Program Studi:

Dr. Bagus Sartono, S.Si., M.Si.

NIP 197804112005011002

Tanggal Ujian:
29 Juli 2026

Tanggal Lulus:



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan November 2025 sampai bulan Agustus 2026 ini ialah *clustering*, dengan judul “*Clustering* Provinsi di Indonesia Berdasarkan Karakteristik Sosial Ekonomi Menggunakan K-Means dan DBSCAN”.

Penulisan karya ilmiah ini tidak lepas dari dukungan dan bimbingan berbagai pihak sehingga karya ilmiah dapat terselesaikan dengan baik. Oleh karenanya, penulis mengucapkan terima kasih kepada:

1. Ayah, Ibu dan kakak-kakak selaku keluarga penulis yang selalu memberikan doa, bimbingan, dukungan dan kasih sayang kepada penulis;
2. Bapak Dr. Septian Rahardiantoro, S.Stat., M.Si. dan Ibu Rahma Anisa S.Stat., M.Si., M.Act.Sc selaku komisi pembimbing yang telah membantu, membimbing, mengarahkan dan mendukung selama penyusunan karya ilmiah ini;
3. Bapak Dr. Bagus Sartono, S.Si., M.Si. selaku moderator kolokium, Bapak Ir. Aam Alamudi, M.Si. selaku moderator seminar hasil, serta Bapak Prof. Dr. Ir. Muhammad Nur Aidi, M.S selaku dosen penguji luar komisi yang telah memberikan saran dan masukan dalam penulisan karya ilmiah ini;
4. Seluruh dosen dan tenaga pendidik Program Studi Statistika dan Sains Data yang telah memberi ilmu yang bermanfaat dan menunjang segala kebutuhan penulis selama perkuliahan dan penyusunan karya ilmiah ini;
5. Adit, Rizam, Rakesh, Goni, Haikal, Firlan, Bintang, Jodi, Oci, Laras, Bulan dan Keyzha selaku sahabat penulis yang telah meluangkan waktu untuk membantu, menemani dan memberi semangat selama penyusunan karya ilmiah ini;
6. Aulia Sukma Tri Rahmawati selaku partner yang telah menemani penulis selama perkuliahan hingga penyusunan karya ilmiah ini;
7. Seluruh pihak yang telah mendukung yang tidak dapat disebutkan satu per satu.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan..

Bogor, Agustus 2026

Muhammad Rizqa Salas



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



DAFTAR ISI

DAFTAR TABEL	xii
DAFTAR GAMBAR	xii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	2
II TINJAUAN PUSTAKA	3
2.1 <i>Variance Inflation factor</i>	3
2.2 Analisis Komponen Utama	3
2.3 Analisis Faktor	4
2.4 <i>Uniform Manifold Approximation and Projection</i>	5
2.5 <i>K-Means Clustering</i>	7
2.6 <i>Density-Based Spatial Clustering of Applications with Noise</i>	8
2.7 Metrik Evaluasi Gerombol	9
2.7.1 <i>Silhouette Index</i>	10
2.7.2 <i>Davies–Bouldin Index</i>	11
2.7.3 <i>Calinski–Harabasz Index</i>	12
III METODE	13
3.1 Data	13
3.2 Prosedur Analisis	14
IV HASIL DAN PEMBAHASAN	17
4.1 Hasil	17
4.1.1 Eksplorasi data dan uji multikolinearitas	17
4.1.2 Reduksi dimensi	18
4.1.3 Penggerombolan dengan K-Means dan DBSCAN	20
4.1.4 Profil gerombol metode terbaik (K-Means x UMAP)	21
4.1.5 Identifikasi provinsi dengan karakteristik ekstrem (<i>outlier</i>)	24
4.2 Pembahasan	25
V SIMPULAN DAN SARAN	29
5.1 Simpulan	29
5.2 Saran	29
DAFTAR PUSTAKA	30
32	
LAMPIRAN	32
RIWAYAT HIDUP	37



DAFTAR TABEL

1	Peubah yang digunakan	14
2	Ringkasan hyperparameter algoritma K-Means dan DBSCAN	16
3	Pasangan peubah dengan korelasi Pearson $ r \geq 0,80$	18
4	Nilai VIF tiap peubah	18
5	Keragaman yang dijelaskan oleh PC1 pada tiap sektor	19
6	Hasil uji kelayakan dan kesesuaian model Analisis Faktor	19
7	Keragaman yang dijelaskan tiap faktor hasil Analisis Faktor (rotasi Varimax)	20
8	Peringkat sepuluh metode <i>clustering</i> berdasarkan <i>Silhouette Index</i>	21
9	Anggota gerombol metode terbaik (K-Means x UMAP, k=3)	22
10	Provinsi yang teridentifikasi sebagai <i>noise/outlier</i> pada tiap metode DBSCAN	25

DAFTAR GAMBAR

1	Ilustrasi skema tahapan UMAP: graf k-NN berbobot, pembentukan <i>fuzzy simplicial set</i> , dan optimasi <i>embedding</i> pada ruang berdimensi rendah (McInnes <i>et al.</i> 2018)	6
2	Ilustrasi skema klasifikasi titik pada DBSCAN: <i>core point</i> , <i>border point</i> , dan <i>noise</i> (Ester <i>et al.</i> 1996)	8
3	Sebaran data terstandarisasi (<i>Z-score</i>) tiap peubah	17
4	Hasil <i>clustering</i> K-Means + UMAP (metode terbaik)	23
5	<i>Radar chart</i> karakteristik rata-rata 17 peubah penelitian antar gerombol (K-Means + UMAP)	24
6	Peta tematik K-Means + UMAP	26
7	Hasil <i>clustering</i> DBSCAN + UMAP	26

DAFTAR LAMPIRAN

1	Rata-rata dan simpangan baku tiap gerombol hasil K-Means + UMAP	33
2	<i>Loading</i> faktor hasil Analisis Faktor (rotasi Varimax)	34
3	Hasil uji sensitivitas nilai <i>n_neighbors</i> pada UMAP	34
4	Peta sebaran gerombol sepuluh metode	35
5	Cluster plot sepuluh metode	36