Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 5, Issue: 2, Month: July, Year: 2026
Page: 719 - 723, Available online at: https://rcf-indonesia.org/home/

PENGELOMPOKAN TINGKAT PENGGUNAAN BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA *K-MEANS CLUSTERING*

Revinta Arrova Dewi
Program Studi Informatika, Universitas Bina Sarana Informatika
Corresponding Author: revintaad10@gmail.com

Article Info

Article history:

Received: Jun 22, 2026
Revised: Jul 10, 2026
Accepted: Jul 19, 2026
Published: Jul 22, 2026

Keywords:

Bahasa daerah
K-Means Clustering
Orange Data Mining
Silhouette Score
Penduduk
Migrasi

ABSTRACT (10 PT)

Indonesia memiliki keragaman bahasa daerah yang sangat tinggi, dengan lebih dari 700 bahasa yang tersebar di berbagai wilayah. Namun, globalisasi, kemajuan teknologi, dan mobilitas penduduk menyebabkan penggunaan bahasa daerah cenderung menurun, khususnya pada generasi muda. Penelitian ini bertujuan mengelompokkan 34 provinsi di Indonesia berdasarkan tingkat penggunaan bahasa daerah menggunakan algoritma *K-Means Clustering*, serta menginterpretasikan hasil pengelompokan menggunakan data migrasi penduduk sebagai variabel pendukung. Data yang digunakan adalah data sekunder Badan Pusat Statistik (BPS) berupa persentase penduduk usia 10 tahun ke atas yang menggunakan bahasa daerah pada tahun 2018, 2021, dan 2024, serta data migrasi masuk, migrasi keluar, dan migrasi netto. Nilai rata-rata ketiga tahun dijadikan variabel utama proses clustering yang dijalankan melalui *Orange Data Mining* dengan jumlah kelompok sebanyak tiga. Kualitas hasil pengelompokan dievaluasi menggunakan *Silhouette Score*. Hasil penelitian menunjukkan terbentuknya kelompok tinggi (12 provinsi), sedang (13 provinsi), dan rendah (9 provinsi) dengan nilai *Silhouette Score* sebesar 0,678 yang mengindikasikan kualitas pengelompokan yang baik. Interpretasi data migrasi menunjukkan bahwa provinsi dengan migrasi masuk tinggi cenderung berada pada kelompok penggunaan bahasa daerah rendah, sedangkan provinsi dengan pengaruh budaya lokal kuat berada pada kelompok tinggi. Hasil penelitian ini dapat menjadi dasar penentuan prioritas kebijakan pelestarian dan revitalisasi bahasa daerah di Indonesia.

1. INTRODUCTION

Indonesia merupakan negara dengan keragaman bahasa daerah yang sangat tinggi [1]. Badan Pengembangan dan Pembinaan Bahasa mencatat terdapat lebih dari 700 bahasa daerah yang digunakan masyarakat di berbagai wilayah Indonesia [2]. Bahasa daerah tidak hanya berfungsi sebagai alat komunikasi, tetapi juga menjadi bagian penting dari identitas budaya dan kearifan lokal, sebagaimana dilindungi melalui Pasal 32 Ayat (2) UUD 1945 yang menegaskan bahwa negara menghormati dan memelihara bahasa daerah sebagai bagian dari kekayaan budaya nasional.

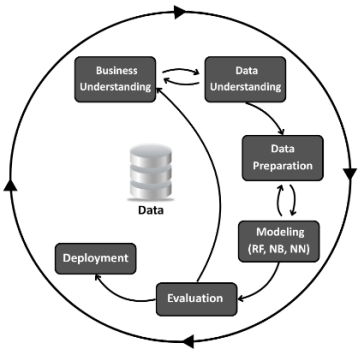
Globalisasi, urbanisasi, dan kemajuan teknologi informasi menyebabkan penggunaan bahasa daerah mengalami penurunan [3]. Hasil *Long Form* Sensus Penduduk 2020 menunjukkan penggunaan bahasa daerah dalam lingkungan keluarga mencapai 74,77%, namun menurun menjadi 72,78% dalam interaksi dengan kerabat dan tetangga, dan menurun lebih tajam pada generasi Z dan generasi Alfa yang hanya berkisar 67-68% [4]. Kondisi ini menimbulkan kekhawatiran terhadap keberlangsungan bahasa daerah [5], terlebih terdapat bahasa daerah yang telah punah maupun mengalami penurunan jumlah penutur secara signifikan.

Mobilitas penduduk turut memengaruhi pola penggunaan bahasa daerah. Provinsi dengan arus migrasi tinggi berpotensi mengalami percampuran bahasa yang lebih besar, sedangkan wilayah dengan kondisi demografis yang relatif homogen umumnya lebih mampu mempertahankan bahasa daerahnya. Oleh karena itu, data migrasi berpotensi menjadi faktor pendukung dalam memahami pola penggunaan bahasa daerah antarprovinsi.

Sejumlah penelitian terdahulu telah membuktikan efektivitas algoritma *K-Means Clustering* pada berbagai ranah sosial, seperti pengelompokan data kemiskinan [6], tingkat pengangguran terbuka [7], Indeks Pembangunan Manusia [8]. Namun, penelitian mengenai pengelompokan tingkat penggunaan bahasa daerah yang dikaitkan dengan data migrasi penduduk masih relatif terbatas, sementara penelitian kebahasaan yang ada umumnya bersifat kualitatif-deskriptif tanpa memanfaatkan teknik *data mining.*

Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan algoritma *K-Means Clustering* untuk mengelompokkan 34 provinsi di Indonesia ke dalam tiga kategori tingkat penggunaan bahasa daerah (tinggi, sedang, rendah), dan menginterpretasikan hasil pengelompokan tersebut menggunakan data migrasi penduduk sebagai variabel pendukung. Kebaruan penelitian ini terletak pada penggabungan analisis bahasa daerah, teknik *data mining*, dan pendekatan demografis berbasis migrasi penduduk, yang diharapkan dapat menjadi dasar penyusunan kebijakan pelestarian bahasa daerah yang lebih tepat sasaran.

2. MATERIALS AND METHODS



Gambar 1 Tahapan Penelitian

Penelitian ini mengacu pada tahapan metode *CRISP-DM (Cross Industry Standard Process for Data Mining)* yang terdiri atas enam tahap utama, yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment.*

2.1. *Business Understanding*

Tahap ini bertujuan memahami konteks permasalahan, yaitu adanya variasi tingkat penggunaan bahasa daerah antarprovinsi di Indonesia yang perlu dipetakan secara objektif berbasis data. Penelitian diarahkan untuk mengelompokkan provinsi ke dalam kategori tinggi, sedang, dan rendah, serta memperkuat interpretasinya menggunakan data migrasi penduduk sebagai konteks demografis pendukung kebijakan pelestarian bahasa daerah.

2.2. *Data Understanding*

Data yang digunakan merupakan data sekunder dari Badan Pusat Statistik (BPS), terdiri atas dua kelompok. Pertama, persentase penduduk usia 10 tahun ke atas yang menggunakan bahasa daerah pada tahun 2018, 2021, dan 2024 untuk 34 provinsi (sebelum pemekaran wilayah). Kedua, data migrasi penduduk per provinsi yang meliputi migrasi masuk, migrasi keluar, dan migrasi netto. Kelompok data pertama digunakan sebagai variabel pembentuk *cluster*, sedangkan data migrasi digunakan sebagai variabel pendukung interpretasi dan tidak dimasukkan ke dalam proses *clustering*. Rincian atribut dataset disajikan pada Tabel 1.

Tabel 1 Atribut Dataset

Atribut	Keterangan	Tipe Data
Provinsi	Nama provinsi di Indonesia (34 provinsi)	Object
Bahasa_ 2018	Persentase penggunaan bahasa daerah tahun 2018	Float
Bahasa_ 2021	Persentase penggunaan bahasa daerah tahun 2021	Float
Bahasa_ 2024	Persentase penggunaan bahasa daerah tahun 2024	Float
Rata-rata	Rata-rata persentase tiga tahun (fitur utama clustering)	Float
Migrasi Masuk	Jumlah migrasi masuk per provinsi (data pendukung)	Integer
Migrasi Keluar	Jumlah migrasi keluar per provinsi (data pendukung)	Integer
Migrasi Netto	Selisih migrasi masuk dan keluar (data pendukung)	Integer

2.3. *Data Preparation*

Tahap pra-proses data diawali dengan menghitung nilai rata-rata persentase penggunaan bahasa daerah tahun 2018, 2021, dan 2024 menggunakan Microsoft Excel, dengan rumus:

$$\bar{x} = (x_1 + x_2 + x_3) / n$$

dengan $\bar{x}$ adalah rata-rata penggunaan bahasa daerah, x_1, x_2, x_3 adalah data tahun 2018, 2021, dan 2024, serta n adalah jumlah tahun. Sebagai contoh, Provinsi Aceh memperoleh nilai rata-rata sebesar 77,10 dari data 76,81; 77,82; dan 76,66. Hasil rata-rata seluruh provinsi ditunjukkan pada Tabel 2 (ringkasan sebagian data).

Tabel 2 Rata-Rata Provinsi

No	PROVINSI	Rata-Rata
1	ACEH	77.10
2	SUMATERA UTARA	42.46
3	SUMATERA BARAT	95.92
4	RIAU	57.09
5	JAMBI	85.07
6	SUMATERA SELATAN	97.23
7	BENGKULU	93.46
8	LAMPUNG	74.33
9	KEP. BANGKA BELITUNG	94.90
10	KEP. RIAU	34.55
11	DKI JAKARTA	5.79
12	JAWA BARAT	73.03
13	JAWA TENGAH	96.57
14	DI YOGYAKARTA	91.36
15	JAWA TIMUR	95.18
16	BANTEN	53.94
17	BALI	87.40
18	NUSA TENGGARA BARAT	90.80
19	NUSA TENGGARA TIMUR	66.10
20	KALIMANTAN BARAT	80.49
21	KALIMANTAN TENGAH	87.86
22	KALIMANTAN SELATAN	95.57
23	KALIMANTAN TIMUR	32.60
24	KALIMANTAN UTARA	24.90
25	SULAWESI UTARA	73.22
26	SULAWESI TENGAH	44.45
27	SULAWESI SELATAN	70.76
28	SULAWESI TENGGARA	46.33
29	GORONTALO	64.59
30	SULAWESI BARAT	73.65
31	MALUKU	70.66

32	MALUKU UTARA	68.33
33	PAPUA BARAT	31.71
34	PAPUA	45.37

Selanjutnya, dataset diimpor ke *Orange Data Mining*. Pemeriksaan *missing values* dilakukan melalui *widget Data Info* dan tidak ditemukan nilai yang hilang, sehingga data dinyatakan lengkap. Pengaturan *tipe atribut* dilakukan melalui *widget Select Columns*, di mana kolom nama provinsi ditetapkan sebagai *Meta Attribute*, kolom rata-rata ditetapkan sebagai satu-satunya *Feature* yang menjadi input utama *K-Means*, sedangkan kolom persentase tahunan dan data migrasi ditempatkan pada bagian *Ignore* karena hanya berfungsi sebagai data pendukung.

2.4. *Modeling*

Proses pengelompokan dilakukan menggunakan algoritma *K-Means Clustering* melalui *widget K-Means* pada *Orange Data Mining*, dengan jumlah *cluster* ditetapkan sebanyak tiga (tinggi, sedang, rendah). Jarak antar data dan *centroid* dihitung menggunakan *Euclidean Distance:*

$$D(x,y) = \sqrt{((x - y)^2)}$$

dengan D adalah jarak, x adalah nilai rata-rata penggunaan bahasa daerah suatu provinsi, dan y adalah nilai *centroid cluster*. Algoritma bekerja secara iteratif: menentukan *centroid* awal, mengelompokkan data berdasarkan jarak terdekat, memperbarui nilai *centroid*, dan mengulang proses hingga *centroid konvergen* (tidak lagi berubah) [9].

2.5. *Evaluation*

Kualitas hasil pengelompokan dievaluasi menggunakan *Silhouette Coefficient* melalui *widget Silhouette Plot* pada *Orange Data Mining*, dengan rumus [10]:

$$S(i) = (b(i) - a(i)) / max\{a(i), b(i)\}$$

dengan S(i) adalah nilai *silhouette* untuk objek i, a(i) adalah rata-rata jarak objek i terhadap objek lain dalam *cluster* yang sama, dan b(i) adalah rata-rata jarak objek i terhadap objek pada *cluster* terdekat yang berbeda. Nilai *silhouette* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan kualitas pengelompokan yang baik [11].

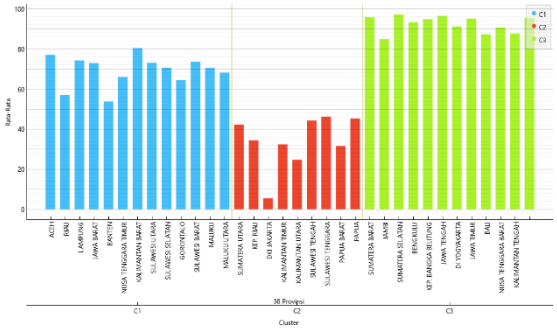
2.6. *Deployment*

Tahap akhir berupa penyusunan hasil evaluasi dan interpretasi *cluster* berdasarkan data migrasi penduduk. Hasil pengelompokan yang diperoleh diharapkan dapat digunakan sebagai bahan pertimbangan bagi Badan Pengembangan dan Pembinaan Bahasa serta pemangku kebijakan terkait dalam menyusun prioritas program pelestarian dan revitalisasi bahasa daerah, khususnya bagi provinsi pada kelompok rendah.

3. RESULTS AND DISCUSSION

3.1. Hasil *Clustering*

Berikut gambar hasil *clustering* pada *widget Bar Plot di Orange Data Mining*:



Gambar 2 Hasil *Clustering*

Proses *clustering* menggunakan *widget K-Means* pada *Orange Data Mining* dengan tiga kelompok berhasil mengelompokkan seluruh 34 provinsi berdasarkan rata-rata persentase penggunaan bahasa daerah. Hasil menunjukkan *Cluster* Tinggi beranggotakan 12 provinsi, *Cluster* Sedang beranggotakan 13 provinsi, dan *Cluster* Rendah beranggotakan 9 provinsi, sebagaimana disajikan pada Tabel 3.

Tabel 3 Hasil Pengelompokan Berdasarkan Kategori

Kategori	**Provinsi Anggota**
Tinggi (12 provinsi)	Sumatera Barat, Sumatera Selatan, Jambi, Bengkulu, Kep. Bangka Belitung, Jawa Tengah, DI Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Barat, Kalimantan Tengah, Kalimantan Selatan
Sedang (13 provinsi)	Aceh, Riau, Lampung, Jawa Barat, Nusa Tenggara Timur, Kalimantan Barat, Sulawesi Utara, Sulawesi Selatan, Sulawesi Barat, Gorontalo, Maluku, Maluku Utara, Banten
Rendah (9 provinsi)	Sumatera Utara, Kep. Riau, DKI Jakarta, Kalimantan Timur, Kalimantan Utara, Sulawesi Tengah, Sulawesi Tenggara, Papua Barat, Papua

3.2. Nilai *Centroid*

Nilai *centroid* akhir tiap *cluster* ditunjukkan pada Tabel 4. *Cluster* dengan centroid tertinggi (92,35) dikategorikan sebagai kelompok Tinggi, *centroid* menengah (69,10) sebagai kelompok Sedang, dan *centroid* terendah (33,13) sebagai kelompok Rendah.

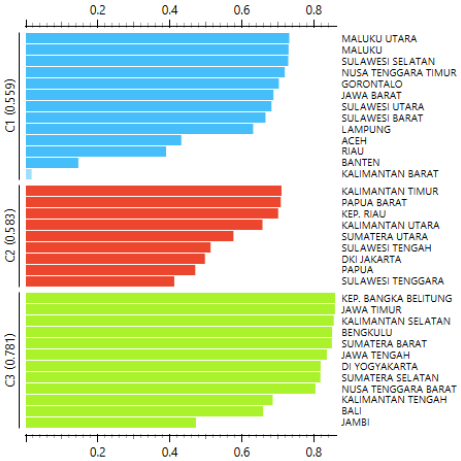
Tabel 4 Nilai *Centeroid* Tiap *Cluster*

Cluster	**Kategori**	**Centroid**
C1	Sedang	69,10
C2	Rendah	33,13
C3	Tinggi	92,35

Sebagai ilustrasi perhitungan jarak manual, Provinsi Aceh dengan rata-rata 77,10 memiliki jarak 8,00 terhadap *centroid* Sedang, 15,25 terhadap *centroid* Tinggi, dan 43,97 terhadap *centroid* Rendah, sehingga Aceh terklasifikasi ke *Cluster* Sedang karena memiliki jarak terkecil. Perhitungan serupa berlaku pada seluruh provinsi lainnya untuk memastikan konsistensi hasil *clustering* dengan keluaran *Orange Data Mining*.

3.3. Evaluasi *Silhouette Score*

Berikut gambar evalusi model pada *Silhoette Plot* di *Orang Data Mining*:



Gambar 3 Hasil Evaluasi Model

Evaluasi kualitas *clustering* menggunakan *widget Silhouette Plot* menghasilkan nilai *Silhouette Score* rata-rata sebesar 0,678. Nilai ini mengindikasikan bahwa pengelompokan yang terbentuk memiliki kualitas yang baik, dengan tingkat kemiripan anggota yang cukup tinggi dalam setiap *cluster*. *Cluster* Tinggi memperoleh nilai *silhouette* tertinggi, menunjukkan tingkat kekompakan anggota yang lebih baik dibandingkan *cluster* lainnya. Sebagai pembanding, perhitungan manual pada Provinsi Aceh tanpa proses normalisasi menghasilkan nilai *silhouette* sebesar 0,431, sedangkan hasil pada *Orange Data Mining* mencapai 0,629 untuk provinsi yang sama. Selisih ini disebabkan oleh proses normalisasi otomatis pada *widget K-Means Orange* yang menyamakan skala data sebelum pengukuran jarak, meskipun kedua metode tetap menghasilkan nilai positif yang mengonfirmasi keanggotaan *cluster* yang sama.

3.4. Interpretasi Berdasarkan Data Migrasi

Untuk memperkuat interpretasi, dilakukan analisis terhadap data migrasi masuk dan migrasi keluar pada setiap kelompok yang terbentuk. Hasil analisis menunjukkan pola yang konsisten: kelompok Rendah cenderung memiliki migrasi masuk yang jauh lebih tinggi dibandingkan kelompok lain, kelompok Sedang berada pada tingkat mobilitas menengah, sedangkan kelompok Tinggi memiliki migrasi masuk yang relatif lebih rendah. Ringkasan interpretasi disajikan pada Tabel 5.

Tabel 5 Interpretasi Hasil *Clustering* Berdasarkan Data Migrasi

Kategori	Karakteristik Migrasi
Tinggi	Migrasi masuk relatif rendah dibandingkan provinsi tujuan migrasi utama; penggunaan bahasa daerah tetap kuat meski terjadi mobilitas penduduk.
Sedang	Migrasi masuk dan keluar bervariasi; mobilitas mulai memengaruhi penggunaan bahasa daerah, namun bahasa daerah masih cukup banyak digunakan.
Rendah	Migrasi masuk tinggi, misalnya DKI Jakarta (3.647.328 jiwa), Kalimantan Timur (1.120.017 jiwa), dan Kepulauan Riau (881.035 jiwa); populasi lebih heterogen sehingga penggunaan bahasa daerah cenderung rendah.

Temuan ini menunjukkan bahwa provinsi yang menjadi tujuan utama migrasi, seperti DKI Jakarta, Kalimantan Timur, dan Kepulauan Riau, cenderung memiliki populasi yang lebih heterogen sehingga penggunaan bahasa daerah menurun. Sebaliknya, provinsi dengan migrasi netto rendah atau negatif serta pengaruh budaya lokal yang kuat, seperti Sumatera Barat, Jawa Tengah, dan Bali, cenderung mempertahankan penggunaan bahasa daerah pada tingkat tinggi meskipun tetap mengalami perpindahan penduduk. Dengan demikian, data migrasi terbukti dapat digunakan sebagai variabel pendukung yang relevan dalam menjelaskan variasi tingkat penggunaan bahasa daerah antarprovinsi, sekaligus menjadi kebaruan penelitian ini dibandingkan penelitian *K-Means* sejenis yang umumnya berhenti pada tahap pengelompokan tanpa interpretasi demografis lanjutan.

4. CONCLUSION

Penelitian ini berhasil menerapkan algoritma *K-Means Clustering* berbasis *Orange Data Mining* untuk mengelompokkan 34 provinsi di Indonesia berdasarkan tingkat penggunaan bahasa daerah, menghasilkan tiga kelompok utama, yaitu kelompok Tinggi (12 provinsi), Sedang (13 provinsi), dan Rendah (9 provinsi), dengan nilai *Silhouette Score* sebesar 0,678 yang menunjukkan kualitas pengelompokan yang baik. Interpretasi menggunakan data migrasi penduduk menunjukkan bahwa provinsi dengan migrasi masuk tinggi cenderung berada pada kelompok penggunaan bahasa daerah rendah, sedangkan provinsi dengan migrasi netto rendah atau negatif dan pengaruh budaya lokal kuat cenderung berada pada kelompok tinggi. Temuan ini mengonfirmasi bahwa pola migrasi dapat digunakan sebagai faktor pendukung yang relevan dalam menginterpretasikan hasil pengelompokan penggunaan bahasa daerah. Penelitian selanjutnya disarankan menambahkan variabel pendukung lain seperti jumlah penutur aktif, tingkat urbanisasi, atau tingkat pendidikan, memperluas cakupan data hingga provinsi hasil pemekaran, serta membandingkan algoritma *K-Means* dengan metode *clustering* lain seperti *K-Medoids* atau *DBSCAN* untuk memperoleh hasil yang lebih komprehensif.

REFERENCES

[1] E. , Luksiawati, N. , Febrina, and F. Iin, "Analisis Ekologi Bahasa Terhadap Faktor Penyebab dan Strategi Pencegahan Kepunahan Bahasa Daerah di Indonesia," *JUPENSAL: Jurnal Pendidikan Universal*, vol. 2, no. 3, pp. 105–117, 2025.

[2] F. Ryandi, "Pelestarian Bahasa Daerah: Tantangan dan Upaya yang Dilakukan Pemerintah," *Info Singkat*, vol. 17, Oct. 2025.

[3] N. B. Rangkuti and E. P. Handayani, "MELESTARIKAN WARISAN LELUHUR: PENGGUNAAN DAN PEMELIHARAAN BAHASA JAWA, MINANG DAN MANDAILING DI ERA MODERN DALAM BAHASA INDONESIA," *Basaya: Jurnal Bahasa, Sastra dan Budaya*, vol. 2, no. 1, pp. 81–90, 2025.

[4] Badan Pusat Statistik, “Profil Suku dan Keberagaman Bahasa Daerah Hasil Long Form Sensus Penduduk 2020,” Jakarta, Dec. 2024.
[5] I. P. R. , Sari, N. I. , Nabila, and R. R. Muhammad, “Ancaman Pergeseran Bahasa Daerah Dan Dampaknya Terhadap Keberlanjutan Warisan Budaya Di Era Global,” *Jurnal Penelitian Nusantara*, vol. 1, no. 5, pp. 91–96, 2025, doi: https://doi.org/10.59435/menulis.v1i5.236.
[6] M. N. , Mursidin, R. A. , Adinda, A. , Komang, and A. D. Risnayanti, “Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Menggunakan Orange Data Mining (Studi Kasus: Kabupaten/Kota Provinsi Sulawesi Selatan),” *Dipanegara Komputer Teknologi Informatika*, vol. 15, no. 2, pp. 277–289, 2022.
[7] A. I. Ramadhan, P. D. Atika, and K. F. Ramdhania, “Analisis Clustering K-Means untuk Pemetaan Tingkat Pengangguran Terbuka di Provinsi-Provinsi Indonesia Tahun 2013-2023,” *Journal of Students ‘Research in Computer Science*, vol. 5, no. 2, pp. 109–122, 2024.
[8] N. N. K. , Kaylista, K. , Nabiilah, V. E. N. F. Gading, and Y. P. Y. Ajif, “Implementasi algoritma K-Means clustering menggunakan aplikasi Orange untuk mengetahui pola indeks pembangunan manusia tahun 2022,” *Journal of Informatic and Information Security*, vol. 4, no. 1, pp. 65–76, 2023.
[9] N. F. , Sutirta and Noviandi, “Perbandingan Manhattan dan Euclidean Distance Untuk Pengelompokan Penyakit Jantung Menggunakan Algoritma K-Means,” Feb. 2024.
[10] A. K. E. , Pily, Susanti., R. , Unang, and Tashid, “Komparasi K-Means Clustering dengan Euclidean dan Cosine Similarity untuk Segmentasi dan Rekomendasi Produk pada Data E-Commerce,” *The Indonesian Journal of Computer Science*, vol. 14, no. 2, Apr. 2025, doi: 10.33022/ijcs.v14i2.4713.
[11] K. J. P. A. Z. , Setiyanto, M. , Amri, and M. H. Kartika, “Improved K-Means Clustering untuk Analisis Ketimpangan Pembangunan Kabupaten/Kota di Pulau Jawa,” 2025.