Jurnal Riset Mahasiswa Matematika
Volume 5 (2) (2025), Pages 123-141
p-ISSN: 2808-1552; e-ISSN: 2808-4926

Segmentasi Provinsi di Indonesia Menggunakan Metode K-Means Berdasarkan Tujuan Mengakses Internet Tahun 2024

Anugrah Putri Nabila and Bunga Mardhotillah*

Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Jambi, Indonesia

Abstrak

Ketimpangan dalam akses digital di Indonesia tidak hanya disebabkan oleh kurangnya infrastruktur, tetapi juga oleh perbedaan cara orang menggunakan internet di tiap daerah. Penelitian ini bertujuan untuk mengelompokkan 38 provinsi di Indonesia berdasarkan tujuan penggunaan internet pada tahun 2024. Metode yang digunakan adalah K-Means Clustering, di mana data dianalisis berdasarkan persentase penduduk yang melakukan tujuh aktivitas digital, seperti ekonomi, pendidikan, dan pekerjaan. Sebelum analisis, dilakukan pra-pemrosesan dengan uji KMO dan VIF untuk memastikan data layak dan tidak ada duplikasi. Dengan metode Elbow, ditentukan ada lima kelompok optimal yang mewakili tipe-tipe penggunaan internet di provinsi. Hasil menunjukkan perbedaan yang jelas, di mana Kelompok 3 menjadi pusat ekonomi digital dengan aktivitas belanja online yang tinggi, sementara Kelompok 1 menunjukkan penggunaan internet yang paling pasif. Meski uji Silhouette Coefficient menunjukkan skor 0,18 yang menandakan kelompok cenderung tumpang tindih, model ini tetap bermanfaat bagi kebijakan. Temuan ini bisa menjadi dasar bagi pemerintah dalam merancang program literasi digital dan kebijakan ekonomi yang lebih sesuai dengan kebutuhan setiap provinsi, sehingga mempercepat transformasi digital yang lebih merata dan inklusif.

Kata Kunci: Ekonomi Digital; K-Means Clustering; Segmentasi Provinsi; Transformasi Digital; Tujuan Akses Internet

Abstract

Digital inequality in Indonesia is no longer solely driven by infrastructure accessibility but also by differences in the functional patterns of internet usage across regions. This study aims to classify 38 provinces in Indonesia based on the characteristics of internet access purposes in 2024. Using the K-Means Clustering method, this research processes population percentage data across seven digital activity variables, including economy, education, and employment. The pre-processing stage involves KMO and VIF tests to ensure data structure adequacy and redundancy minimization. Based on the Elbow method, five optimal clusters were determined to represent the digital typology of the provinces. The analysis results show a clear stratification, where Cluster 3 stands out as a digital economy hub with high e-commerce activity, while Cluster 1 reflects the most passive digital involvement. Although the Silhouette Coefficient test yielded a score of 0.18, indicating a tendency for overlapping cluster structures, the model still provides significant policy utility. These findings offer a strategic basis for the government to design digital literacy interventions and economic policies that are specifically tailored to the unique needs of each provincial group, aiming to accelerate a more inclusive national digital transformation.

Keywords: Digital Economy; Digital Transformation; Internet Access Purpose; K-Means Clustering; Provincial Segmentation

*Corresponding author. E-mail: bunga.mstat08@unja.ac.id

Submitted: 14 November 2025 Reviewed: 21 Desember 2025 Accepted: 30 Desember 2025
DOI: https://dx.doi.org/10.18860/jrmm.v5i2.38985

1 Pendahuluan

Bidang analisis data geografis dan perilaku digital semakin penting dalam memahami perbedaan dan potensi pembangunan daerah di Indonesia. Akses dan penggunaan internet menjadi indikator penting dalam menilai tingkat modernisasi dan kesetaraan digital, yang mencerminkan tidak hanya ketersediaan infrastruktur tetapi juga bagaimana masyarakat mengadopsi teknologi [1], [2]. Memahami perbedaan pola penggunaan internet di antara provinsi sangat penting untuk merancang kebijakan pembangunan infrastruktur, penguatan ekonomi digital, dan program literasi digital yang lebih tepat sasaran. Penelitian ini berfokus pada pola penggunaan internet berdasarkan tujuan spesifik di tingkat provinsi, yang merefleksikan aktivitas masyarakat dalam aspek ekonomi, sosial, pendidikan, dan pekerjaan di era digital.

Penelitian sebelumnya memang sering membahas faktor-faktor yang memengaruhi penggunaan internet, seperti tingkat pendidikan dan kondisi sosial-ekonomi. Namun, banyak penelitian masih menggunakan data agregat atau fokus pada tingkat penetrasi internet secara umum, sehingga belum cukup menjelaskan dimensi fungsional dari pemanfaatan internet. Beberapa kajian deskriptif mengenai kesenjangan digital di Indonesia menunjukkan adanya perbedaan antardaerah dan tantangan literasi digital yang masih menonjol [2], sementara kajian terkait *digital divide* juga menekankan pentingnya melihat variasi pemanfaatan TIK secara spasial [1]. Di sisi lain, metode *unsupervised learning* seperti klastering (clustering) dipandang efektif untuk mengidentifikasi pola tersembunyi dalam data multidimensi, dan K-Means merupakan salah satu metode yang banyak digunakan untuk keperluan pengelompokan data numerik [3], [4]. Penentuan jumlah klaster yang optimal umumnya dilakukan dengan Metode Elbow [5], sedangkan kualitas pemisahan klaster dapat dievaluasi menggunakan *silhouette coefficient* [6].

Kesenjangan riset utama dalam literatur saat ini adalah kecenderungan penelitian terdahulu yang lebih banyak memotret "siapa" yang memiliki akses internet (aksesibilitas) atau "seberapa besar" penetrasi internet, namun belum cukup menekankan "untuk apa" internet digunakan (fungsionalitas) di tingkat kewilayahan. Padahal, tujuan penggunaan internet (misalnya untuk informasi/berita, media sosial, jual-beli, belajar daring, bekerja dari rumah, hiburan, dan produksi konten digital) dapat mencerminkan kesiapan ekonomi digital dan kapasitas literasi digital suatu daerah. Selain itu, karena K-Means berbasis jarak Euclidean, standarisasi data diperlukan agar variabel memiliki kontribusi yang seimbang; salah satu pendekatan yang lazim digunakan adalah Z-score [7], [8]. Aspek lain yang perlu dicermati adalah sensitivitas K-Means terhadap nilai ekstrem, sehingga deteksi outlier menjadi langkah penting untuk menjaga stabilitas hasil pengelompokan [9].

Penelitian ini bertujuan untuk mengklasifikasikan provinsi di Indonesia berdasarkan data persentase penduduk yang menggunakan internet menurut tujuan akses internet pada tahun 2024. Metode yang digunakan adalah K-Means Clustering, untuk mengidentifikasi kelompok-kelompok provinsi yang memiliki pola penggunaan internet yang serupa [3], [4]. Kebaruan penelitian ini terletak pada integrasi tujuh dimensi tujuan penggunaan internet sebagai basis segmentasi dengan menggunakan data terbaru tahun 2024, sehingga mampu menangkap dinamika perilaku digital masyarakat yang relevan dengan penguatan ekonomi digital, pembelajaran daring, dan kerja jarak jauh.

Pendekatan ini memungkinkan identifikasi klaster yang lebih spesifik dibandingkan pembagian provinsi berdasarkan penetrasi internet atau kriteria geografis semata. Kontribusi penelitian ini adalah dua aspek: secara ilmiah, hasil klastering memperkaya literatur mengenai geografi digital di Indonesia melalui klasifikasi perilaku penggunaan internet; secara praktis, profil klaster dapat digunakan sebagai dasar bagi pemerintah pusat maupun daerah dalam merancang strategi intervensi digital yang lebih tepat sasaran, termasuk penguatan ekosistem ekonomi digital dan peningkatan literasi digital sesuai kebutuhan tiap kelompok provinsi.

Artikel ini terdiri dari struktur berikut. Bagian 2 membahas metode termasuk langkah-langkah pengolahan dan pemodelan data menggunakan K-Means, penentuan jumlah klaster

dengan Metode Elbow [5], serta evaluasi kualitas klaster dengan *silhouette coefficient* [6]. Bagian 3 berisi hasil dan pembahasan analisis. Bagian 4 menyampaikan kesimpulan beserta saran untuk penelitian lebih lanjut.

2 Metode

Dalam bab ini dijelaskan secara lengkap metode yang digunakan dalam penelitian, mulai dari tahap perencanaan penelitian, pengumpulan dan persiapan data, sampai cara melakukan analisis clustering. Subbab berikutnya membahas tentang desain penelitian sebagai langkah awal untuk memahami pendekatan dan proses yang digunakan dalam membagi geografis (provinsi) menjadi kelompok menggunakan metode K-Means clustering.

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan analitis deskriptif dan komputasional unsupervised learning untuk melakukan segmentasi geografis (provinsi) dengan metode K-Means clustering. Tujuannya adalah mengelompokkan provinsi-provinsi di Indonesia yang memiliki pola perilaku digital serupa, khususnya dalam hal tujuan mengakses internet, berdasarkan data persentase penduduk. K-Means dipilih karena mampu mengelompokkan data angka secara efisien dan bisa menemukan pola tersembunyi tanpa perlu label awal. Penelitian ini memiliki desain deskriptif-analitik, di mana hasil pemisahan provinsi akan digunakan untuk memberikan deskripsi mendalam mengenai disparitas dan fokus digital antar-regional, yang menjadi dasar rekomendasi kebijakan digital yang lebih tepat sasaran. Penelitian ini dirancang agar prosedurnya dapat direplikasi dengan data BPS periode berikutnya.

2.2 Sumber Data dan Persiapan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari Modul Statistik Telekomunikasi Indonesia Tahun 2024 yang dipublikasikan oleh Badan Pusat Statistik (BPS). Data ini mencakup persentase penduduk di setiap provinsi di Indonesia yang menggunakan internet untuk berbagai tujuan spesifik pada tahun 2024. Data mentah melalui serangkaian proses pra-pemrosesan data untuk memastikan kualitas dan kelayakan komputasi. Proses ini meliputi pemeriksaan kelengkapan data, validasi, deteksi outlier, dan standarisasi data. Setelah itu dilakukan uji asumsi untuk memastikan metode statistik atau algoritma yang digunakan pada metode K-Means clustering valid dan menghasilkan kesimpulan yang dapat diandalan. Adapun uji asumsi yang dilakukan adalah pengecekkan sampel representatif menggunakan Uji Kaiser Meyer Olkin (KMO) dan pengecekkan asumsi multikolinearitas menggunakan nilai Variance Inflation Factor (VIF).

2.3 Variabel Penelitian

Dalam penelitian ini, penulis menggunakan tujuh variabel yang diambil dari data persentase penduduk yang mengakses internet menurut provinsi dan tujuan mengakses internet tahun 2024.

Tabel 1: Variabel penelitian

Variabel	Definisi	Tipe Data
Info/Berita	Persentase penduduk yang mengakses internet untuk mencari informasi umum atau berita	Numerik (Persentase)
Media Sosial	Persentase penduduk yang mengakses internet untuk menggunakan platform media sosial	Numerik (Persentase)
Jual Barang/Jasa	Persentase penduduk yang mengakses internet untuk aktivitas menjual barang atau jasa (e-commerce)	Numerik (Persentase)
Belajar Online	Persentase penduduk yang mengakses internet untuk kegiatan pembelajaran daring	Numerik (Persentase)
WFH	Persentase penduduk yang mengakses internet untuk keperluan bekerja dari rumah (Work From Home)	Numerik (Persentase)
Hiburan	Persentase penduduk yang mengakses internet untuk tujuan hiburan (misalnya streaming film, musik)	Numerik (Persentase)
Konten Digital	Persentase penduduk yang mengakses internet untuk membuat atau mengunggah konten digital	Numerik (Persentase)

Ketujuh variabel pada Tabel 1 selanjutnya akan digunakan dalam proses analisis K-Means Clustering untuk mengidentifikasi pola dan pengelompokan geografis (provinsi).

2.4 Pra-Pemrosesan Data

Tahap awal dalam proses pra-pemrosesan data adalah mendeteksi dan mengatasi berbagai permasalahan dalam dataset, seperti nilai yang hilang, data ganda, maupun data yang tidak sesuai atau tidak relevan [3]. Proses ini dimulai dengan mengidentifikasi missing value atau nilai yang tidak tersedia pada variabel tertentu yang dapat menghambat komputasi algoritma. Selanjutnya, dilakukan pemeriksaan terhadap data duplikat untuk memastikan tidak ada unit amatan yang tercatat lebih dari satu kali sehingga pengulangan informasi yang sama dapat dihindari. Selain itu, peneliti juga melakukan seleksi terhadap data yang tidak relevan atau data yang tidak memberikan kontribusi langsung terhadap tujuan pemetaan perilaku digital provinsi.

Proses ini dimulai dengan mengidentifikasi missing value atau nilai yang tidak tersedia pada variabel tertentu yang dapat menghambat komputasi algoritma. Selanjutnya, dilakukan pemeriksaan terhadap data duplikat untuk memastikan tidak ada unit amatan yang tercatat lebih dari satu kali sehingga pengulangan informasi yang sama dapat dihindari. Selain itu, peneliti juga melakukan seleksi terhadap data yang tidak relevan atau data yang tidak memberikan kontribusi langsung terhadap tujuan pemetaan perilaku digital provinsi.

2.5 Standarisasi Data

Standarisasi dilakukan untuk menyamakan skala variabel numerik agar tidak ada variabel yang mendominasi analisis clustering. Konsep Z-score merupakan salah satu metode normalisasi data yang digunakan untuk melakukan standarisasi terhadap nilai-nilai dalam dataset [7]. Metode Z-Score bertujusn untuk menyamakan skala variabel numerik dan dirumuskan pada Pers. 1 [10][8].

$$X' = \frac{X - \bar{X}}{S} \tag{1}$$

dengan,
X' = nilai hasil standarisasi
X = nilai asli
$\bar{X}$= rata-rata variabel
S = standar deviasi variabel

Dengan standar deviasi = 1 dan mean = 0, semua variabel memiliki kontribusi yang seimbang pada perhitungan jarak Euclidean dalam K-Means.

2.6 Deteksi Outlier

Outlier didefinisikan sebagai data dengan nilai ekstrem yang berada jauh dari distribusi umum dataset secara keseluruhan. Keberadaan outlier merupakan aspek krusial dalam analisis data karena dapat memengaruhi performa algoritma clustering, terutama K-Means yang memiliki sensitivitas tinggi terhadap jarak antar titik data. Pendeteksian outlier secara teoretis dilakukan untuk meminimalisasi dampak noise serta meningkatkan kualitas dan validitas hasil pengelompokan yang dihasilkan [9].

Dalam literatur statistika dan analisis data, identifikasi pencilan umumnya dilakukan melalui beberapa pendekatan standar. Pendekatan tersebut meliputi penggunaan boxplot untuk mendeteksi titik data yang berada di luar jangkauan whisker, perhitungan Z-Score untuk mengidentifikasi nilai yang melampaui ambang batas standar deviasi tertentu, serta metode Interquartile Range (IQR) yang menetapkan batas pencilan pada data yang berada di luar rentang pagar luar bawah atau atas. Secara konseptual, penanganan terhadap outlier yang terdeteksi dapat dilakukan melalui beberapa strategi, yakni penghapusan data untuk meningkatkan stabilitas model, transformasi data untuk menekan besaran nilai ekstrem, atau tetap mempertahankan nilai tersebut jika dianggap membawa informasi penting yang relevan dengan konteks fenomena yang diteliti.

2.7 Pemeriksaan Karakteristik Data (Pra-Klatering)

Meskipun algoritma K-Means tidak membutuhkan asumsi statistik yang begitu ketat seperti pada model regresi, tetap perlu dilihat ciri-ciri data agar hasil pemisahan datanya tetap bagus. Langkah ini tidak jadi syarat utama agar algoritma bisa berjalan, tetapi untuk memahami cara data itu mengelompok dan mengurangi informasi yang berlebihan yang bisa mengganggu perhitungan jarak antar data. Pada tahap ini dilakukan penilaian kelayakan struktur data dengan uji KMO dan identifikasi redundansi antar variabel independent dengan nilai VIF [11].

2.7.1 Kecukupan Struktur Data (KMO)

Uji Kaiser-Meyer-Olkin (KMO) dalam penelitian ini digunakan sebagai tahap pengecekan tambahan untuk menilai apakah struktur data sudah layak untuk dilakukan proses klastering. Nilai KMO mengukur cukup tidaknya sampel dengan melihat hubungan antar variabel. Alasan mengapa KMO digunakan dalam konteks K-Means adalah untuk memastikan bahwa variabel-variabel yang menggambarkan akses internet memiliki pola keterkaitan yang cukup kuat, sehingga dapat dijadikan dasar untuk dibentuk menjadi kelompok-kelompok (klaster) yang bermakna. Pengujian kelayakan data untuk analisis klaster dilakukan menggunakan Kaiser Meyer Olkin (KMO) yang dirumuskan pada Pers. 2.

$$\text{KMO} = \frac{\sum_{k \neq l} r_{kl}^2}{\sum_{k \neq l} r_{kl}^2 + \sum_{k \neq l} p_{kl}^2} \tag{2}$$

dengan,

$$r_{X_k X_l} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ik} - \bar{X}_k\right)\left(X_{il} - \bar{X}_l\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ik} - \bar{X}_k\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{il} - \bar{X}_l\right)^2}}$$

$$\rho_{X_k X_l | X_m} = \frac{r_{X_k X_l} - r_{X_k X_m} r_{X_l X_m}}{\sqrt{\left(1 - r_{X_k X_m}^2\right)\left(1 - r_{X_l X_m}^2\right)}}$$

Keterangan:
ρ = banyaknya variabel
n = banyaknya objek
$r_{X_k X_l}$ = korelasi antara variabel X_k dan X_l
$\overline{X}_k$ = rata-rata variabel X_k
$\overline{X}_l$ = rata-rata variabel X_l
$\rho_{X_k X_l \cdot X_m}$ = korelasi parsial antara variabel X_k dan X_l dengan menjaga agar X_m konstan.
Jika nilai KMO yang diperoleh dari data antara 0,5 sampai 1 maka menunjukkan bahwa dataset memiliki struktur yang baik dan siap untuk diproses lebih lanjut.

2.7.2 Identifikasi Redundansi Variabel (VIF)

Variance Inflation Factor (VIF) digunakan untuk mengetahui apakah terdapat korelasi yang terlalu tinggi antar variabel independen. Dalam algoritma K-Means yang menggunakan jarak Euclidean, adanya dua atau lebih variabel dengan korelasi sangat tinggi dapat menyebabkan pengaruh yang terlalu besar secara tidak langsung pada dimensi tersebut. Jika nilai VIF melebihi batas (VIF $>$ 10), maka ini menunjukkan adanya variabel yang memberikan informasi yang hampir sama. Identifikasi adanya multikolinearitas antar variabel dilakukan menggunakan Variance Inflation Factor (VIF) sebagaimana ditunjukkan pada Pers. 3.

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{3}$$

Keterangan:
$[VIF]_i$ = nilai Variance Inflation Factor ke-i
R_i^2 = koefisien determinasi yang diperoleh bila nilai variabel ke $-$ i yang diregresikan dengan variabel lainnya
i = variabel dengan i = 1, 2, ..., p
Pemeriksaan VIF dilakukan agar pastikan setiap variabel yang digunakan memberikan informasi yang berbeda dan tidak saling bergantung dalam menentukan jarak antar provinsi.

2.8 Penentuan Jumlah Cluster Optimal

Metode Elbow digunakan untuk menentukan jumlah cluster yang optimal dalam penerapan algoritma clustering, seperti pada metode K-Means [5]. Prinsipnya adalah memplot Within-Cluster Sum of Squares (WCSS) terhadap jumlah cluster, dan memilih titik "elbow" di mana penurunan WCSS mulai melambat. Penentuan jumlah cluster optimal dalam metode K-Means dilakukan berdasarkan nilai Within Cluster Sum of Squares (WCSS) yang dirumuskan pada Pers. 4 [12].

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4}$$

Keterangan:
C_i = cluster ke-i
μ_i = centroid cluster ke-i
$\|x - \mu_i\|$ = jarak Euclidean
Elbow membantu menentukan k yang efisien, tidak terlalu kecil (kurang representatif) atau terlalu besar (overfitting).

2.9 Kualitas Cluster

Silhouette Score adalah cara untuk mengetahui seberapa baik suatu objek berada dalam kelompok tertentu dibandingkan dengan kelompok lain. Nilai silhouette untuk satu objek i dirumuskan

pada Pers. 5 [6]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \tag{5}$$

Keterangan:

a(i) = rata-rata jarak antara objek i dengan semua objek lain di dalam cluster yang sama

b(i) = rata-rata jarak antara objek i dan objek terdekat dari cluster lain (cluster yang berbeda).

Nilai $s(i)$ berkisar dari -1 hingga $+1$. Nilai mendekati $+1$ menunjukkan objek sangat sesuai dengan clusternya dan jauh dari cluster lain; nilai mendekati 0 menunjukkan objek berada di batas antar cluster; sedangkan nilai negatif menunjukkan objek mungkin salah ditempatkan dalam cluster.

2.10 Algoritma K-Means

K-Means Clustering merupakan salah satu metode klasterisasi non-hirarki yang berfungsi untuk mengelompokkan data ke dalam satu atau beberapa klaster berdasarkan kesamaan karakteristiknya [4]. Prosedur kerja algoritma ini dilakukan secara iteratif yang dimulai dengan menentukan jumlah klaster (k) yang akan dibentuk, kemudian dilanjutkan dengan menetapkan pusat awal klaster atau centroid secara acak. Langkah selanjutnya adalah menghitung jarak setiap titik data ke masing-masing centroid menggunakan ukuran jarak tertentu, di mana dalam banyak implementasi digunakan jarak Euclidean. Data kemudian dikelompokkan ke dalam klaster dengan centroid terdekat, diikuti dengan pembaruan posisi centroid berdasarkan nilai rata-rata dari seluruh titik data dalam klaster tersebut. Seluruh proses ini diulangi secara terus-menerus hingga posisi centroid mencapai titik konvergen atau tidak mengalami perubahan lagi.Penghitungan jarak antara objek dengan centroid memegang peranan krusial dalam menentukan keanggotaan klaster. Perhitungan jarak Euclidean yang umum digunakan dalam algoritma K-Means ditunjukkan pada Pers. 6 [13][14].

$$d(x_i, \mu_j) = \sqrt{\sum_{p=1}^{n} (x_{ip} - \mu_{jp})^2} \tag{6}$$

Keterangan:

$d(x_i, \mu_j)$ = jarak antara data ke-i dan centroid ke-j

x_{ip} = nilai data ke-i pada variabel ke-p

μ_{jp} = nilai centroid ke-j pada variabel ke-p

n = jumlah variabel

Dalam konteks analisis geografis, penerapan K-Means memberikan deskripsi mendalam mengenai disparitas dan fokus digital antar-regional, yang menjadi dasar rekomendasi kebijakan digital yang lebih tepat sasaran.

2.11 Perangkat Lunak

Penelitian ini menggunakan perangkat lunak Python melalui platform Google Colab. Google Colab digunakan sebagai lingkungan kerja berbasis cloud untuk menjalankan seluruh proses analisis data, mulai dari membersihkan data, memperbaiki standar, mendeteksi nilai yang tidak normal, melakukan uji asumsi untuk K-Means, menentukan jumlah kluster yang paling sesuai dengan metode Elbow, menghitung nilai Silhouette Score, hingga menerapkan algoritma K-Means clustering. Beberapa library seperti pandas, numpy, scikit-learn, dan matplotlib digunakan untuk membantu mengolah data, melakukan perhitungan angka, dan menampilkan hasil secara visual.. Dengan menggunakan perangkat lunak ini, proses analisis bisa berjalan efektif, tepat, dan terorganisir.

2.12 Alur Kerja Penelitian

Alur Kerja Alur kerja dalam penelitian ini disusun secara sistematis untuk memastikan proses analisis berjalan terstruktur dari tahap awal hingga penarikan kesimpulan. Tahapan dimulai dengan pengumpulan data melalui proses pengunduhan dan impor dataset, yang kemudian dilanjutkan dengan tahap pra-pemrosesan data untuk membersihkan nilai yang hilang serta memastikan relevansi data yang akan digunakan. Mengingat variabel dalam penelitian ini memiliki satuan yang mungkin berbeda, dilakukan standarisasi data menggunakan StandardScaler agar seluruh variabel memiliki skala yang seragam sebelum memasuki proses klasterisasi. Selanjutnya, dilakukan deteksi outlier dengan metode statistik seperti Interquartile Range (IQR) guna mengidentifikasi nilai-nilai ekstrem yang berpotensi memengaruhi stabilitas hasil pengelompokan.

Sebelum algoritma diimplementasikan, dilakukan pemeriksaan karakteristik data dengan uji KMO untuk melihat kelayakan struktur data dan pemeriksaan multikolinearitas dengan melihat nilai VIF untuk memastikan tidak ada redundansi antar variabel independen. Tahap krusial berikutnya adalah penentuan jumlah klaster optimal melalui metode Elbow berdasarkan nilai Within-Cluster Sum of Square (WCSS). Kualitas dari hasil pengelompokan tersebut kemudian dievaluasi menggunakan Silhouette Score untuk menilai konsistensi dan pemisahan antar klaster. Setelah jumlah klaster yang paling sesuai ditetapkan, algoritma K-Means diimplementasikan secara penuh pada dataset. Alur penelitian ini diakhiri dengan visualisasi dan interpretasi hasil melalui analisis grafik serta tabel, yang kemudian menjadi dasar bagi penarikan kesimpulan untuk menghubungkan temuan klasterisasi dengan tujuan penelitian yang telah ditetapkan.

3 Hasil dan Pembahasan

Pada bagian ini, disajikan secara rinci temuan dan interpretasi dari proses segmentasi provinsi menggunakan K-Means Clustering. Berdasarkan data multivariat tahun 2024, analisis ini mengelompokkan 38 provinsi di Indonesia menjadi klaster-klaster yang berbeda, yang masing-masing merepresentasikan profil penggunaan internet yang unik, mulai dari fokus pada informasi, sosial, hingga ekonomi digital.

3.1 Data Penelitian

Data yang menjadi basis utama penelitian ini adalah data sekunder multivariat yang diterbitkan oleh Badan Pusat Statistik (BPS) melalui Modul Statistik Telekomunikasi Indonesia Tahun 2024. Data ini mencakup 38 entitas provinsi di Indonesia. Fokus analisis berada pada tujuh variabel krusial yang merefleksikan prioritas penggunaan internet di tingkat regional pada tahun 2024: persentase penduduk yang menggunakan internet untuk Info/Berita, Media Sosial, Jual Barang/Jasa, Belajar Online, WFH, Hiburan, dan Konten Digital. Ragam dimensi ini menyediakan fondasi yang kuat untuk mengidentifikasi pola segmentasi provinsi menggunakan K-Means Clustering.

Tabel 2: Data penelitian

Provinsi	Tujuan Mengakses Internet						
	Info/Berita (%)	Media Sosial (%)	Jual Barang/Jasa (%)	Belajar Online (%)	WFH (%)	Hiburan (%)	Konten Digital (%)
Aceh	80.37	77.00	2.26	7.53	1.64	82.91	3.98
Sumatera Utara	76.81	76.18	2.86	8.04	1.02	86.22	4.50
Sumatera Barat	76.55	78.62	3.24	9.38	1.31	88.31	5.25
Riau	77.32	76.55	3.27	9.19	1.13	89.05	6.30
Jambi	75.89	76.34	4.46	9.46	1.28	86.56	10.60
Sumatera Selatan	74.77	77.56	4.02	9.75	1.15	86.08	10.98
Bengkulu	76.05	81.23	5.06	8.19	1.22	84.82	9.21
Lampung	71.43	75.98	5.60	9.52	1.20	84.55	7.94
Kep. Bangka Belitung	77.88	96.13	5.11	8.92	1.20	91.89	7.51
Kep. Riau	80.24	79.07	6.23	10.41	2.05	89.53	1.70
DKI Jakarta	83.80	80.56	7.21	11.21	3.40	83.05	9.81
Jawa Barat	77.24	76.72	5.32	11.41	1.86	83.96	4.74
Jawa Tengah	76.31	80.13	6.19	9.10	1.09	85.92	10.84
DI Yogyakarta	77.67	82.93	9.97	10.52	2.65	82.89	14.45
Jawa Timur	78.91	76.03	5.88	10.51	1.60	84.20	7.24
Banten	80.46	74.14	5.96	10.99	2.29	82.63	2.96
Bali	85.78	84.87	6.87	10.61	1.51	90.22	11.86
NTB	68.41	70.12	4.37	9.54	1.52	91.87	10.50
NTT	76.54	75.39	2.42	8.79	1.11	86.59	8.28
Kalimantan Barat	76.53	77.92	3.06	7.48	1.06	86.24	2.74
Kalimantan Tengah	78.36	76.06	4.65	7.88	1.33	87.56	4.49
Kalimantan Selatan	75.25	78.99	4.87	9.12	1.37	88.93	11.89
Kalimantan Timur	78.15	79.43	5.02	8.55	1.31	86.32	3.95
Kalimantan Utara	72.04	78.46	4.13	9.07	1.02	83.26	2.84
Sulawesi Utara	78.25	82.75	5.95	10.23	2.09	84.37	10.00
Sulawesi Tengah	74.63	76.06	4.08	8.87	1.47	86.16	2.44
Sulawesi Selatan	77.99	78.25	4.36	12.21	1.65	85.10	7.77
Sulawesi Tenggara	77.33	78.42	3.57	8.64	0.88	89.49	10.14
Gorontalo	76.21	80.71	4.38	7.48	1.13	84.44	5.32
Sulawesi Barat	72.49	72.10	2.46	8.38	1.29	85.27	6.71
Maluku	80.45	80.33	2.13	7.78	2.31	87.42	6.96
Maluku Utara	67.81	70.94	2.24	6.96	1.18	79.51	1.51
Papua Barat	72.65	69.33	1.70	6.38	1.27	85.00	1.68
Papua Barat Daya	73.27	75.56	3.12	6.32	1.42	80.73	0.90
Papua	81.32	77.26	2.27	9.24	2.45	86.20	2.03
Papua Selatan	79.27	76.22	4.03	6.47	1.17	88.75	14.32
Papua Tengah	71.77	66.99	2.82	5.27	0.52	71.19	7.36
Papua Pegunungan	75.22	71.36	2.79	3.04	0.79	73.93	0.40

Data pada Tabel 2 belum dapat langsung digunakan untuk proses analisis dan harus melalui tahap pra-pemrosesan terlebih dahulu agar siap digunakan secara optimal dalam analisis lebih lanjut.

3.2 Pra-Pemrosesan Data

Sebelum dilakukan proses Clustering, langkah awal yang penting adalah memastikan bahwa data memenuhi syarat penggunaan metode K-Means Clustering.

3.2.1 Pengecekkan Missing Value

Pemeriksaan dilakukan dengan menghitung jumlah nilai kosong pada setiap variabel. Berikut hasil pemeriksaan missing value pada persentase penduduk yang mengakses internet menurut provinsi dan tujuan mengakses internet tahun 2024:

Tabel 3: Pengecekkan missing value

Variabel	Jumlah Missing Value
Provinsi	0
Info/Berita (%)	0
Media Sosial (%)	0
Jual Barang/Jasa (%)	0
Belajar Online (%)	0
WFH (%)	0
Hiburan (%)	0
Konten Digital (%)	0

Dari Tabel 3, terlihat bahwa seluruh variabel memiliki jumlah Missing Value sebesar nol.

3.2.2 Pengecekkan Data Duplikat

Pengecekkan duplikat dilakukan untuk memastikan bahwa setiap data provinsi hanya dihitung satu kali untuk menghindari bias dalam analsis yang disebabkan oleh data ganda. Berikut hasil pemeriksaan data duplikat: Jumlah data duplikat: np.int64(0) Berdasarkan hasil pemeriksaan, tidak ditemukan adanya data duplikat di seluruh baris data. Hal ini menginformasikan bahwa bahwa setiap entri provinsi dalam himpunan data adalah unik dan tidak ada baris yang merepresentasikan provinsi yang sama secara berulang. Sehingga data siap digunakan untuk analisis klastering tanpa risiko bias dari pengamatan ganda.

3.2.3 Pengecekkan Data Relevan

Pemeriksaan relevan data bertujuan untuk memastikan bahwa data yang digunakan dalam proses analisis benar-benar sesuai dengan konteks penelitian dan tidak mengandung nilai yang dapat menyebabkan hasil analisis menjadi bias. Hasil pengecekan relevan data sebagai berikut:

Tabel 4: Pengecekkan relevan data

Kategori Pengecekan	Hasil Pengecekan
Cek kelengkapan data provinsi	Jumlah baris unik provinsi: 38
Cek validitas nilai persentase semua variabel (0%–100%)	Valid
Cek tipe data variabel tujuan internet	Numerik (float64)

Berdasarkan hasil yang diperoleh pada Tabel 4, seluruh pemeriksaan menghasilkan data yang relevan karena jumlah provinsi sudah memadai untuk analisis regional, semua variabel berada pada rentang (0-100) persen, dan semua variabel sudah bertipe numerik . Dengan demikian, data dinyatakan layak dari segi kelengkapan, validitas, dan konsistensi tipe data untuk melanjutkan ke tahap analisis multivariat.

3.3 Standarisasi Data

Proses standarisasi menggunakan teknik standarisasi StandarScaler yang dilakukan dengan tujuan untuk menyamakan skala antar variabel numerik sehingga tidak ada satu variabel yang mendominasi variabel lainnya dalam proses analisis Clustering. Setelah dilakukan proses standarisasi, hasil perubahan nilai masing-masing variabel adalah sebagai berikut:

Tabel 5: Data setelah proses standarisasi

Provinsi	Tujuan Mengakses Internet						
	Info/Berita (%)	Media Sosial (%)	Jual Barang/Jasa (%)	Belajar Online (%)	WFH (%)	Hiburan (%)	Konten Digital (%)
Aceh	1.021	-0.084	-1.202	-0.683	0.293	-0.585	-0.697
Sumatera Utara	0.052	-0.253	-0.851	-0.397	-0.818	0.228	-0.561
Sumatera Barat	-0.018	0.248	-0.628	0.354	-0.298	0.741	-0.363
Riau	0.191	-0.177	-0.611	0.247	-0.621	0.923	-0.087
Jambi	-0.197	-0.220	0.085	0.398	-0.352	0.311	1.042
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Papua Barat Daya	-0.910	-0.381	-0.699	-1.362	-0.101	-1.121	-1.507
Papua	1.279	-0.031	-1.197	0.275	1.747	0.223	-1.210
Papua Selatan	0.722	-0.450	-0.166	-1.289	-0.549	0.845	2.020
Papua Tengah	-1.318	-2.143	-0.874	-1.950	-1.716	-3.467	0.190
Papua Pegunungan	-0.379	-1.244	-0.892	-3.201	-1.231	-2.793	-1.639

Dengan standarisasi ini, terlihat pada Tabel 5 seluruh variabel berada pada skala yang sebanding, sehingga perhitungan jarak Euclidean menjadi lebih adil dan hasil pengelompokan provinsi dapat lebih objektif, tidak terpengaruh oleh perbedaan satuan atau rentang nilai antar variabel.

3.4 Deteksi Outlier

Identifikasi outlier dilakukan agar kita bisa memahami data ekstrem yang mungkin memengaruhi penentuan pusat massa atau centroid dalam algoritma K-Means. Dalam penelitian ini, deteksi outlier dilakukan dengan melihat seluruh unit observasi provinsi secara menyeluruh melalui analisis multivariat, bukan hanya melihat temuan outlier pada setiap variabel secara terpisah. Pendekatan ini dipilih agar tidak terjadi kesalahan penghitungan yang sama dua kali, karena hal itu bisa mengganggu pemahaman kita tentang sifat data secara keseluruhan. Berikut hasil deteksi Outlier pada setiap variabel:

Tabel 6: Deteksi outlier

Aktivitas	Jumlah Outlier
Info/Berita (%)	4
Media Sosial (%)	5
Jual Barang/Jasa (%)	1
Belajar Online (%)	1
WFH (%)	3
Hiburan (%)	2
Konten Digital (%)	0

Berdasarkan Tabel 6, ditemukan beberapa provinsi memiliki nilai ekstrem pada variabel tujuan akses internet tertentu, namun data tersebut tetap digunakan dalam analisis. Keputusan untuk mempertahankan nilai-nilai tersebut didasarkan pada pertimbangan bahwa data tersebut mencerminkan keragaman nyata dalam akses digital di berbagai wilayah Indonesia. Jika data tersebut dihilangkan, informasi mengenai provinsi-provinsi yang mengalami akselerasi digital sangat tinggi dibanding wilayah lainnya akan terlewat. Meski demikian, diketahui bahwa metode K-Means standar sensitif terhadap nilai ekstrem karena menggunakan jarak Euclidean yang memperkuat perbedaan deviasi. Hal ini membuat nilai-nilai pencilan dapat memengaruhi posisi pusat klaster. Oleh karena itu, penelitian ini mengakui bahwa metode yang digunakan belum sepenuhnya robust terhadap nilai-nilai ekstrem, sehingga interpretasi terhadap klaster yang

meliputi provinsi-provinsi ekstrem dilakukan dengan mempertimbangkan dampak nilai-nilai tersebut terhadap posisi pusat klaster akhir.

3.5 Hasil Pemeriksaan Karakteristik Data

Sebelum mengimplementasikan algoritma K-Means, dilakukan pemeriksaan karakteristik data untuk memastikan bahwa variabel-variabel yang digunakan memiliki kualitas yang baik dan tidak mengandung redundansi informasi yang ekstrem. Langkah ini penting untuk menjamin bahwa hasil pengelompokan mencerminkan pola data yang sebenarnya tanpa distorsi dari variabel yang saling tumpang tindih secara linier.

3.5.1 Hasil Pemeriksaan Struktur Data (KMO)

Untuk melihat kelayakan struktur data, dapat dilihat pada nilai Kaiser Mayer Olkin (KMO):

Hipotesis:

H_0 = Dataset memiliki struktur yang baik untuk diproses lebih lanjut

H_1 = Dataset belum memiliki struktur yang baik untuk diproses lebih lanjut

Hasil pengujian:

Nilai KMO Keseluruhan: np. float64 (0,6503093110934051)

Berdasarkan hasil analisis diperoleh nilai KMO secara keseluruhan $\geq 0,5$, hal ini berarti data memiliki kecukupan struktur yang memadai (clustering tendency). Nilai KMO yang berada di atas ambang batas 0,5 memberikan konfirmasi bahwa variabel-variabel dalam dataset memiliki keterkaitan pola yang cukup kuat untuk dikelompokkan ke dalam klaster-klaster yang bermakna melalui metode K-Means.

3.5.2 Hasil Pemeriksaan Redundansi Variabel (VIF)

Pemeriksaan multikolinearitas dilakukan dengan melihat nilai Variance Inflation Factor (VIF) untuk memastikan tidak ada redundansi antar variabel independen yang dapat mendistorsi perhitungan jarak Euclidean. Berikut hasil nilai VIF untuk masing-masing variabel:

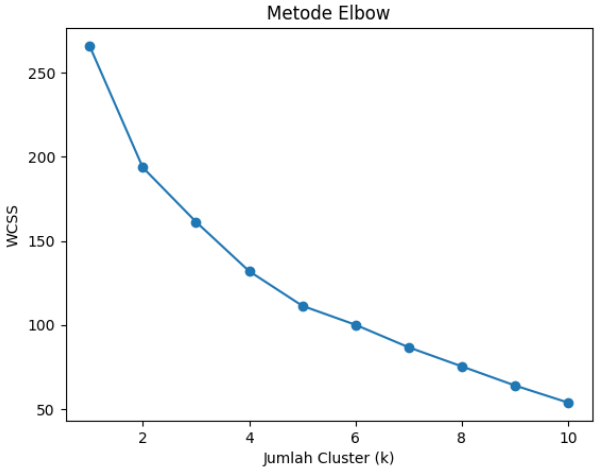
Tabel 7: Nilai vif setiap variabel

Variabel	VIF
Info/Berita (%)	1.866875
Media Sosial (%)	2.088592
Jual Barang/Jasa (%)	2.695218
Belajar Online (%)	2.598811
WFH (%)	2.111903
Hiburan (%)	1.963559
Konten Digital (%)	1.555459

Berdasarkan Tabel 7, semua variabel di bawah 10. Hal ini mengindikasikan bahwa setiap variabel memberikan kontribusi informasi yang unik dan tidak terdapat korelasi antar variabel yang sangat ekstrem. Dengan terpenuhinya kondisi ini, pembentukan klaster dapat dilakukan secara objektif karena setiap dimensi variabel memiliki bobot kontribusi yang seimbang dalam algoritma K-Means.

3.6 Penentuan Jumlah Cluster Optimal

Metode yang digunakan untuk menentukan jumlah cluster optimal adalah Metode Elbow, dimana jumlah kelompok ditetapkan berdasarkan grafik penurunan nilai jumlah kuadrat dalam kelompok (WCSS). Berikut hasil cluster menurut metode elbow:



Gambar 1: Grafik metode elbow

Pada Gambar 1, bisa dilihat bahwa titik elbow muncul pada nilai k = 5, di mana penurunan WCSS mulai melambat. Untuk memberikan kejelasan atas hasil segmentasi, berikut disajikan nilai WCSS untuk setiap jumlah cluster yang diterapkan, dari k = 1 sampai k = 10:

Tabel 8: Nilai wcss tiap cluster

k	WCSS
1	266.000000
2	193.689895
3	161.361722
4	131.361722
5	111.373112
6	100.073519
7	86.679849
8	75.316379
9	63.952548
10	53.767002

Berdasarkan data yang ada pada Tabel 8, nilai WCSS menunjukkan pengurangan yang signifikan sampai k = 5, dan penurunan setelahnya cenderung lebih kecil. Oleh karena itu, jumlah cluster yang paling sesuai menurut Metode Elbow adalah 5, karena pada titik ini terlihat adanya pergeseran yang paling mencolok.

3.7 Kualiatas Cluster

Penentuan jumlah klaster (k) merupakan langkah penting untuk memastikan hasil segmentasi yang sesuai dan bisa digunakan dalam praktik. Dalam penelitian ini, nilai k ditentukan dengan menggabungkan metode Elbow dan Silhouette Coefficient. Dari hasil metode Elbow pada Gambar [Nomor Gambar], terlihat bahwa penurunan jumlah kesalahan dalam klaster (WCSS) mulai menurun secara stabil pada titik $k = 5$, yang menunjukkan bahwa menambah jumlah klaster setelah titik tersebut tidak lagi memberikan pengurangan variasi yang signifikan. Namun, dari pengujian Silhouette Coefficient, diperoleh nilai rata-rata sebesar 0,18. Secara teknis, nilai ini menunjukkan struktur klaster yang kurang kuat dan adanya kemungkinan tumpang tindih antar klaster di beberapa wilayah provinsi.

Meskipun nilai Silhouette Coefficient tergolong rendah, nilai $k = 5$ dipilih sebagai kesepakatan terbaik setelah mempertimbangkan beberapa opsi jumlah klaster lain seperti $k = 4$ dan $k = 6$. Pada $k = 4$, karakteristik wilayah terlalu umum sehingga tidak bisa membedakan daerah dengan

profil ekonomi digital yang berbeda. Di sisi lain, penggunaan $k = 6$ menghasilkan klaster dengan jumlah anggota sangat sedikit, yang kurang efektif untuk mengambil keputusan kebijakan nasional. Secara substansial, nilai $k = 5$ mampu menunjukkan perbedaan pergerakan akses internet di Indonesia secara lebih rinci, mulai dari daerah dengan akses internet pasif hingga pusat perkembangan ekonomi digital. Nilai Silhouette yang rendah (0,18) ini dipahami sebagai gambaran nyata dari penyebaran akses internet di Indonesia yang memang memiliki kesamaan antar wilayah yang berdekatan, sehingga pemisahan secara tegas sulit dicapai tanpa kehilangan informasi penting. Dengan demikian, $k = 5$ tetap dipertahankan karena memiliki manfaat lebih tinggi dalam pembuatan kebijakan dibandingkan model yang memiliki hasil statistik lebih kuat namun kurang informatif secara praktis.

3.8 Hasil Clustering

Mengikuti penentuan jumlah klaster optimal ($K = 5$) yang diperoleh melalui visualisasi Metode Elbow, analisis dilanjutkan dengan menjalankan proses K-Means Clustering. Implementasi ini sukses mengelompokkan data seluruh provinsi ke dalam lima klaster regional yang terpisah. Lima klaster yang dihasilkan merepresentasikan profil adaptasi dan prioritas akses internet yang beragam, di mana setiap klaster (K0 hingga K4) akan dibahas secara rinci berdasarkan nilai rata-rata variabelnya untuk membedakan karakteristik dominan masing-masing segmen.

Tabel 9: Jumlah anggota tiap cluster

Cluster	Jumlah Data
0	14
1	5
2	9
3	2
4	8

Jumlah pada Tabel 9 menunjukkan bahwa Cluster 0 merupakan kelompok dengan anggota terbanyak, sedangkan Cluster 3 merupakan kelompok dengan anggota paling sedikit. Adapun rata-rata untuk kelima cluster adalah sebagai berikut:

Tabel 10: Rata-rata tiap variabel per cluster

Cluster	Info/Berita (%)	Media Sosial (%)	Jual Barang/Jasa (%)	Belajar Online (%)	WFH (%)	Hiburan (%)	Konten Digital (%)
0	76.737857	78.638571	3.505000	8.382857	1.320714	86.545714	5.090714
1	72.144000	70.836000	2.534000	5.594000	1.036000	78.072000	2.370000
2	74.967778	77.110000	4.685556	8.864444	1.208889	87.438889	10.713333
3	80.735000	81.745000	8.590000	10.865000	3.025000	82.970000	12.130000
4	80.023750	78.636250	5.355000	10.701250	1.937500	85.776250	6.037500

Dari Tabel 10 terlihat bahwa setiap cluster memiliki titik pusat (centroid) yang mencerminkan karakteristik khas dari masing-masing kelompok provinsi. Berikut adalah data persentase penduduk yang mengakses internet menurut provinsi dan tujuan mengakses internet tahun 2024 yang telah dicluster:

Tabel 11: Data hasil cluster

Provinsi	Tujuan Mengakses Internet							Cluster
	Info/Berita (%)	Media Sosial (%)	Jual Barang/Jasa (%)	Belajar Online (%)	WFH (%)	Hiburan (%)	Konten Digital (%)	
Aceh	80.37	77.00	2.26	7.53	1.64	82.91	3.98	0
Sumatera Utara	76.81	76.18	2.86	8.04	1.02	86.22	4.50	0
Sumatera Barat	76.55	78.62	3.24	9.38	1.31	88.31	5.25	0
Riau	77.32	76.55	3.27	9.19	1.13	89.05	6.30	0
Jambi	75.89	76.34	4.46	9.46	1.28	86.56	10.60	2
Sumatera Selatan	74.77	77.56	4.02	9.75	1.15	86.08	10.98	2
Bengkulu	76.05	81.23	5.06	8.19	1.22	84.82	9.21	2
Lampung	71,43	75.98	5.60	9.52	1.20	84.55	7.94	2
Kep. Bangka Belitung	77.88	96.13	5.11	8.92	1.20	91.89	7.51	0
Kep. Riau	80.24	79.07	6.23	10.41	2.05	89.53	1.70	4
DKI Jakarta	83.80	80.56	7.21	11.21	3.40	83.05	9.81	3
Jawa Barat	77.24	76.72	5.32	11.41	1.86	83.96	4.74	4
Jawa Tengah	76.31	80.13	6.19	9.10	1.09	85.92	10.84	2
DI Yogyakarta	77.67	82.93	9.97	10.52	2.65	82.89	14.45	3
Jawa Timur	78.91	76.03	5.88	10.51	1.60	84.20	7.24	4
Banten	80.46	74.14	5.96	10.99	2.29	82.63	2.96	4
Bali	85.78	84.87	6.87	10.61	1.51	90.22	11.86	4
Nusa Tenggara Barat	68.41	70.12	4.37	9.54	1.52	91.87	10.50	2
Nusa Tenggara Timur	76.54	75.39	2.42	8.79	1.10	86.59	8.28	0
Kalimantan Barat	76.53	77.92	3.06	7.48	1.06	86.24	2.74	0
Kalimantan Tengah	78.36	76.06	4.65	7.88	1.33	87.56	4.49	0
Kalimantan Selatan	75.25	78.99	4.87	9.13	1.37	88.93	11.89	2
Kalimantan Timur	78.15	79.43	5.02	8.55	1.31	86.32	3.95	0
Kalimantan Utara	72.04	78.46	4.13	9.07	1.02	83.26	2.84	0
Sulawesi Utara	78.25	82.75	5.95	10.23	2.09	84.37	10.00	4
Sulawesi Tengah	74.63	76.06	4.08	8.87	1.47	86.16	2.44	0
Sulawesi Selatan	77.99	78.25	4.36	12.21	1.65	85.10	7.77	4
Sulawesi Tenggara	77.33	78.42	3.57	8.64	0.88	89.49	10.14	2
Gorontalo	76.21	80.71	4.38	7.48	1.30	84.44	5.32	0
Sulawesi Barat	72.49	72.10	2.46	8.38	1.29	85.27	6.71	0
Maluku	80.45	80.33	2.13	7.80	2.31	87.42	6.96	0
Maluku Utara	67.81	70.94	2.24	6.96	1.18	79.51	1.51	1
Papua Barat	72.65	69.33	1.70	6.38	1.27	85.50	1.68	1
Papua Barat Daya	73.27	75.56	3.12	6.32	1.42	80.73	0.90	1
Papua	81.32	77.26	2.27	9.24	2.45	86.20	2.03	4
Papua Selatan	79.27	75.22	4.03	6.45	1.17	88.73	14.32	2
Papua Tengah	71.77	66.99	2.82	5.27	0.52	71.19	7.36	1
Papua Pegunungan	75.22	71.36	2.79	3.04	0.79	73.93	0.40	1

Berdasarkan Tabel 11, terlihat bahwa provinsi-provinsi di Indonesia terbagi ke dalam beberapa cluster berdasarkan kemiripan tujuan dalam mengakses internet. Setiap cluster menunjukkan pola berbeda yang mencerminkan variasi kebutuhan serta perilaku digital masyarakat antarprovinsi. Hasil pengelompokan ini dapat menjadi dasar dalam perumusan kebijakan pengembangan teknologi informasi yang lebih tepat sasaran.

3.9 Interpretasi Hasil Clustering

Interpretasi klaster dilakukan dengan menganalisis nilai rata-rata centroid yang disajikan dalam Tabel 10, yang mencerminkan profil perilaku digital dominan dari provinsi-provinsi yang tergabung di setiap klaster.

Cluster 0 dicirikan oleh pola penggunaan internet yang relatif tinggi pada aktivitas hiburan, meskipun tidak menjadi yang tertinggi dibandingkan klaster lain. Pada klaster ini, pemanfaatan internet untuk tujuan produktif masih terbatas, tercermin dari rendahnya persentase penggunaan untuk jual barang atau jasa, bekerja dari rumah, serta aktivitas belajar daring. Penggunaan konten digital juga berada pada tingkat rendah–menengah, menunjukkan bahwa internet lebih banyak dimanfaatkan sebagai sarana konsumsi daripada penciptaan nilai ekonomi atau peningkatan kapasitas. Kondisi ini mengindikasikan perlunya intervensi kebijakan yang berfokus pada peningkatan literasi digital dan dorongan pemanfaatan internet secara lebih produktif, khususnya untuk kegiatan ekonomi dan pengembangan keterampilan.

Cluster 1 merupakan klaster dengan tingkat pemanfaatan internet terendah di hampir seluruh variabel yang diamati, baik untuk kebutuhan informasi, sosial, hiburan, maupun aktivitas produktif. Rendahnya nilai rata-rata pada jual barang atau jasa, bekerja dari rumah, dan konten digital menunjukkan keterbatasan pemanfaatan internet dalam mendukung aktivitas ekonomi dan pekerjaan. Pola ini mencerminkan adanya kesenjangan digital yang signifikan, yang dapat berkaitan dengan keterbatasan akses, kualitas jaringan, maupun kapabilitas pengguna. Oleh karena itu, strategi intervensi pada klaster ini perlu diprioritaskan pada perluasan infrastruktur

dasar, peningkatan kualitas layanan internet, serta penguatan literasi digital tingkat dasar agar masyarakat mampu memanfaatkan internet untuk fungsi-fungsi fundamental.

Cluster 2 menunjukkan intensitas penggunaan internet yang sangat tinggi untuk hiburan dan disertai dengan tingkat aktivitas konten digital yang relatif tinggi dibandingkan sebagian besar klaster lain. Meskipun demikian, pemanfaatan internet untuk aktivitas ekonomi seperti jual barang atau jasa dan bekerja dari rumah masih tergolong rendah. Pola ini mengindikasikan bahwa klaster ini berada pada fase transisi, di mana penggunaan internet tidak lagi semata-mata bersifat konsumtif, tetapi mulai mengarah pada aktivitas kreasi konten, meskipun belum sepenuhnya terkonversi menjadi kegiatan ekonomi digital. Intervensi kebijakan yang tepat untuk klaster ini adalah penguatan ekosistem kreator digital, termasuk pelatihan produksi konten, dukungan monetisasi, serta peningkatan literasi keamanan dan perlindungan hak cipta.

Cluster 3 dicirikan oleh tingkat pemanfaatan internet yang tinggi dan merata di hampir seluruh variabel, khususnya pada aktivitas produktif seperti jual barang atau jasa, belajar daring, bekerja dari rumah, dan konten digital. Selain itu, penggunaan internet untuk memperoleh informasi dan berinteraksi melalui media sosial juga berada pada tingkat yang tinggi. Pola ini menunjukkan bahwa klaster ini telah memiliki ekosistem digital yang relatif matang, di mana internet berfungsi sebagai infrastruktur utama dalam mendukung aktivitas ekonomi, pendidikan, dan pekerjaan. Strategi kebijakan pada klaster ini sebaiknya difokuskan pada peningkatan kualitas dan kapasitas layanan digital, serta pengembangan sumber daya manusia berkeahlian lanjut untuk menjaga keberlanjutan dan daya saing ekonomi digital.

Cluster 4 memperlihatkan pola pemanfaatan internet yang relatif seimbang dengan penekanan kuat pada aktivitas belajar daring, yang nilainya hampir setara dengan klaster paling produktif. Aktivitas ekonomi digital dan bekerja dari rumah berada pada tingkat menengah, lebih tinggi dibandingkan klaster dengan pemanfaatan rendah, namun belum mencapai intensitas tertinggi. Hal ini menunjukkan bahwa klaster ini berada pada tahap kesiapan digital yang cukup baik, khususnya dalam pengembangan kapasitas sumber daya manusia, tetapi masih memiliki ruang untuk peningkatan produktivitas ekonomi digital. Oleh karena itu, intervensi kebijakan pada klaster ini dapat diarahkan pada akselerasi pemanfaatan keterampilan digital yang telah dimiliki menjadi aktivitas ekonomi nyata, seperti penguatan UMKM digital dan perluasan peluang kerja jarak jauh.

Hasil segmentasi menjadi lima klaster ini telah berhasil mengidentifikasi pola perilaku digital yang berbeda antar-provinsi, mulai dari fokus pada konsumsi hiburan hingga pusat produktivitas ekonomi digital. Profil klaster yang spesifik ini memberikan dasar empiris yang kuat untuk merumuskan kebijakan intervensi digital yang tepat sasaran di tingkat regional.

3.10 Keterbatasan Penelitian dan Saran Lanjutan

Penelitian ini berhasil memetakan 38 provinsi di Indonesia berdasarkan tujuan penggunaan internet pada tahun 2024. Namun, peneliti menyadari bahwa ada beberapa keterbatasan yang perlu diperhatikan demi menjaga integritas hasil penelitian. Dalam evaluasi kualitas klaster, jumlah klaster yang dipilih yaitu $k = 5$ didasarkan pada metode Elbow. Hasilnya, rata-rata Silhouette Score sebesar 0,18. Secara statistik, nilai ini menunjukkan bahwa struktur klaster cenderung lemah dan terdapat tumpang tindih antar wilayah. Meskipun demikian, peneliti tetap memilih model ini karena memiliki interpretasi kebijakan yang paling informatif dalam menggambarkan perbedaan perilaku digital di Indonesia. Meski demikian, nilai silhouette yang rendah ini diakui sebagai keterbatasan teknis, yang menunjukkan bahwa perbedaan antar beberapa provinsi di batas klaster tidak terlalu jelas. Dalam keputusan ini, peneliti mempertahankan data pencilan agar tetap bisa merepresentasikan keragaman penggunaan internet yang nyata di Indonesia. Sayangnya, hal ini berpotensi memengaruhi akurasi posisi centroid klaster karena algoritma K-Means sangat rentan terhadap nilai ekstrem melalui perhitungan jarak Euclidean.

Keterbatasan lainnya adalah variabel yang digunakan masih hanya fokus pada tujuan akses internet tanpa mempertimbangkan faktor pendukung eksternal seperti kualitas infrastruktur fisik

dan indikator sosial ekonomi. Selain itu, penggunaan data cross-section hanya selama satu tahun belum mampu menangkap perubahan perilaku digital secara jangka panjang. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan data multi-tahun agar bisa memetakan tren pergeseran perilaku digital dari waktu ke waktu. Penelitian juga bisa dikembangkan dengan memasukkan variabel makro seperti infrastruktur jaringan dan PDRB per kapita untuk meneliti hubungan antara fasilitas fisik dengan penggunaan internet masyarakat. Selain itu, penggunaan algoritma klastering yang lebih kuat terhadap data pencilan dan tumpang tindih, seperti K-Medoids atau Fuzzy C-Means, sangat dianjurkan untuk memvalidasi konsistensi pengelompokan. Terakhir, penelitian masa depan dapat diperluas ke tingkat kabupaten atau kota untuk menghasilkan gambaran intervensi kebijakan yang lebih detail dan tepat sasaran di tingkat lokal.

4 Kesimpulan

Penelitian ini berhasil memetakan bagaimana 38 provinsi di Indonesia digolongkan berdasarkan tujuan mengapa orang menggunakan internet pada tahun 2024. Dengan menerapkan algoritma K-Means, penelitian ini menjawab kebutuhan untuk membagi wilayah secara lebih detail dibandingkan hanya menilai seberapa luas internet digunakan. Hasil utama menunjukkan adanya lima kelompok provinsi dengan karakteristik penggunaan internet yang berbeda, mulai dari daerah yang menjadi pusat aktivitas ekonomi digital hingga wilayah yang penggunaannya masih sifatnya pasif. Temuan ini menunjukkan bahwa strategi transformasi digital di tingkat nasional harus disesuaikan dengan karakteristik setiap kelompok agar kebijakan bisa lebih tepat sasaran.

Kekuatan dari penelitian ini terletak pada penggunaan tujuh dimensi variabel yang spesifik mengenai tujuan akses internet serta penggunaan dataset terbaru setelah transformasi digital secara besar-besaran di Indonesia. Hal ini memberikan kontribusi penting bagi bidang geografi digital dengan menyediakan klasifikasi perilaku pengguna internet yang lebih dinamis. Secara praktis, profil kelompok ini bisa digunakan oleh pihak berwenang untuk menentukan prioritas program, seperti memperkuat infrastruktur ekonomi kreatif di wilayah dengan aktivitas digital tinggi atau meningkatkan literasi dasar di wilayah yang masih tertinggal.

Meskipun memberikan gambaran perencanaan yang berharga, penelitian ini masih memiliki keterbatasan. Struktur kelompok yang terbentuk belum sangat berbeda secara statistik (skor silhouette 0,18) serta algoritma masih pe sensitif terhadap data yang tidak biasa. Sebagai arah penelitian berikutnya, disarankan untuk menggabungkan variabel infrastruktur fisik dan indikator sosial ekonomi makro agar analisis lebih lengkap. Selain itu, penggunaan metode pengelompokan yang lebih stabil terhadap data yang tumpang tindih serta penelitian berdasarkan waktu (multi-year) sangat dianjurkan agar bisa memantau perubahan tren penggunaan internet di tiap provinsi di Indonesia secara lebih rinci.

Pernyataan Kontribusi Penulis (CRediT)

Anugrah Putri Nabila: Konseptualisasi, Metodologi, Perangkat Lunak, Kurasi Data, Analisis Formal, Visualisasi, Validasi, Penulisan Draf Awal. **Bunga Mardhotillah:** : Supervisi, Administrasi Proyek, Penulisan Telaah dan Penyuntingan.

Deklarasi Penggunaan AI atau Teknologi Berbasis AI

Gemini digunakan untuk membantu penyusunan draf awal, perbaikan struktur kalimat, pembuatan serta penyesuaian syntax pemrograman, dan membantu proses penanganan error selama analisis.

Deklarasi Konflik Kepentingan

Artikel ini disusun sebagai bagian dari pemenuhan persyaratan Program MBKM Studi Independen.

Pendanaan dan Ucapan Terima Kasih

Penelitian ini mendapatkan bantuan dana untuk biaya menerbitkan artikel dari dosen pembimbing, yang juga memberikan panduan, pengawasan, dan saran penting selama proses penelitian dan penulisan. Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Badan Pusat Statistik (BPS) Republik Indonesia, khususnya tim penyusun Modul Statistik Telekomunikasi Indonesia Tahun 2024, atas ketersediaan data publik yang memungkinkan terlaksananya analisis segmentasi provinsi ini. Ucapan terima kasih juga diberikan kepada pihak-pihak lain yang membantu atau memberikan masukan penting dalam menyelesaikan penelitian, meskipun tidak tercantum sebagai penulis.

Ketersediaan Data

Data yang digunakan dalam penelitian ini, yaitu persentase penduduk yang mengakses internet berdasarkan tujuan di tingkat provinsi pada tahun 2024, bersumber dari Modul Statistik Telekomunikasi Indonesia Tahun 2024 yang dipublikasikan oleh Badan Pusat Statistik (BPS) Republik Indonesia. Data mentah yang menjadi dasar analisis utama, telah disajikan secara rinci dalam bagian Hasil dan Pembahasan artikel ini. Oleh karena itu, data yang relevan untuk replikasi dan verifikasi temuan penelitian ini bersifat terbuka dan tersedia untuk publik melalui penerbitan resmi BPS.

Daftar Pustaka

[1] H. D. Ramadhanti and E. T. Astuti, "Digital divide and a spatial investigation of convergence in ict development across provinces in indonesia," *Journal of Regional Development*, pp. 69–84, 2023.

[2] I. Widiasanti, S. Rahmadani, and D. A. Nur, "Kesetaraan akses internet dan tantangan literasi digital di indonesia," *Jurnal Ilmiah Nasional*, vol. 9, pp. 19 631–19 637, 2025.

[3] A. F. Zabidi, "Penerapan algoritma k-means untuk pengelompokan koleksi perpustakaan dengan data mining," *Media Jurnal Informatika*, vol. 16, no. 2, p. 233, 2024. DOI: 10.35194/mji.v16i2.4814

[4] R. A. Indraputra and R. Fitriana, "K-means clustering data covid-19," *Jurnal Teknik Informatika*, vol. 10, no. 3, pp. 275–282, 2020.

[5] N. T. Hartanti, "Metode elbow dan k-means guna mengukur kesiapan siswa smk dalam ujian nasional," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 6, no. 2, pp. 82–89, 2020. DOI: 10.25077/teknosi.v6i2.2020.82-89

[6] A. Kassambara, *Practical Guide to Cluster Analysis in R*. STHDA, 2015.

[7] I. W. Pratama and P. Putra, "Standarisasi z-score sebagai pendekatan alternatif dalam evaluasi prestasi akademik mahasiswa," *Jurnal Pendidikan Tinggi*, vol. 1, no. 2, pp. 77–85, 2023.

[8] D. Abdullah, *Statistika Terapannya pada Bidang Informatika*. Graha Ilmu, 2014.

[9] P. A. Ariawan, "Optimasi pengelompokan data pada metode k-means dengan analisis outlier," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 5, no. 2, pp. 88–95, 2019. DOI: 10.25077/teknosi.v5i2.2019.88-95

[10] F. N. A. Wahidah Nilam, J. Oktalia, and O. Juwita, "Pengelompokkan daerah rawan bencana di indonesia menggunakan metode clustering k-means," *Jupiter: Jurnal Ilmu Keteknikan Industri Teknik Elektro dan Informatika*, vol. 3, no. 1, pp. 9–16, 2024. DOI: 10.61132/jupiter.v3i1.644

[11] T. W. I. R. Fahrur and I. T. Utami, "Implementasi algoritma k-medoids dan k-error untuk pengelompokan kabupaten/kota di provinsi jawa tengah berdasarkan jumlah produksi peternakan tahun 2020," *Jurnal Gaussian*, vol. 11, pp. 366–376, 2023. DOI: 10.14710/j.gauss.11.3.366-376

[12] E. Prayitno, I. J. Perdana, E. Iskandar, B. Heri, and A. A. Subagyo, "Optimalisasi profitabilitas ritel melalui segmentasi pelanggan dengan k-means clustering," *Jurnal Manajemen Bisnis*, vol. 9, no. 3, pp. 113–120, 2024.

[13] F. N. Cahya, Y. Mahatma, and S. R. Rohimah, "Perbandingan metode perhitungan jarak euclidean dan manhattan pada k-means clustering," *JMT: Jurnal Matematika dan Terapan*, vol. 5, no. 1, pp. 43–55, 2023. DOI: 10.21009/jmt.5.1.5

[14] M. Kurniawan, *Kupas Tuntas Algoritma Clustering*. Informatika Bandung, 2021.