

KLASTERISASI KABUPATEN/KOTA DI PROVINSI JAWA TENGAH BERDASARKAN KOMPONEN INDEKS PEMBANGUNAN MANUSIA MENGGUNAKAN METODE UMAP DAN K-MEANS**Difta Alzena Sakhi**

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur

e-mail: 23083010061@student.upnjatim.ac.id***Friza Nur Fatmala**

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur

e-mail: 23083010051@student.upnjatim.ac.id**Karina Auralia**

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur

e-mail: 23083010072@student.upnjatim.ac.id**Alfan Rizaldy Pratama**

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur

e-mail: alfan.fasilkom@upnjatim.ac.id**Aviolla Terza Damaliana**

Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur

e-mail: aviolla.terza.sada@upnjatim.ac.id**Abstrak**

Ketimpangan pembangunan manusia di Provinsi Jawa Tengah tetap menjadi tantangan besar yang memerlukan pendekatan berbasis data. Penelitian ini mengelompokkan kabupaten/kota berdasarkan komponen Indeks Pembangunan Manusia (IPM) tahun 2019–2024 menggunakan algoritma K-Means. Variabel yang dianalisis meliputi Usia Harapan Hidup, Harapan Lama Sekolah, Rata-Rata Lama Sekolah, dan Pengeluaran per Kapita. Data diproses melalui tahapan penting, seperti standarisasi, deteksi outlier, dan reduksi dimensi menggunakan *Uniform Manifold Approximation and Projection* (UMAP), serta penentuan jumlah kluster optimal dengan *Elbow Method* dan *Silhouette Score*. Hasil analisis menunjukkan empat kluster optimal dengan nilai *Silhouette Score* sebesar 0,71, yang mengelompokkan data tahunan seluruh kabupaten/kota ke dalam kelompok-kelompok dengan tingkat pembangunan manusia yang berbeda secara signifikan. Kluster dengan nilai IPM tertinggi terdiri dari kota-kota besar, seperti Semarang, Salatiga, dan Surakarta, sementara kluster terendah mencakup wilayah yang masih menghadapi berbagai kendala dalam aspek kesehatan, pendidikan, dan ekonomi. Visualisasi UMAP membantu interpretasi distribusi kluster dan memberikan masukan strategis bagi kebijakan pembangunan wilayah yang lebih merata.

Kata Kunci: klasterisasi, indeks pembangunan manusia, K-Means, UMAP, Jawa Tengah

Abstract

Human development disparities in Central Java Province remain a major challenge that requires a data-driven approach. This study categorizes districts/cities based on the components of the Human Development Index (HDI) for the years 2019–2024 using the K-Means algorithm. The variables analyzed include Life Expectancy, Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditure. The data was processed through important stages, such as standardization, outlier detection, and dimension reduction using *Uniform Manifold Approximation and Projection* (UMAP), as well as determining the optimal number of clusters using the Elbow method and *Silhouette Score*. The analysis results showed four optimal clusters with a *Silhouette Score* of 0.71, grouping the annual data of all districts/cities into clusters with significantly different levels of human development. The cluster with the highest HDI consists of large cities, such as Semarang, Salatiga, and Surakarta, while the lowest cluster includes areas that still face various challenges in health, education, and economic aspects. UMAP visualization aids in interpreting cluster distribution and provides strategic input for more equitable regional development policies.

Keywords: clustering, human development index, K-Means, UMAP, Central Java

PENDAHULUAN

Pembangunan berkelanjutan telah menjadi fokus utama bagi banyak negara dalam upaya meningkatkan kesejahteraan masyarakat secara menyeluruh. Pada tahun 1990, United Nations Development Programme (UNDP) memperkenalkan konsep baru untuk mengukur pembangunan melalui Indeks Pembangunan Manusia (IPM) yang mengintegrasikan aspek kesehatan, pendidikan, dan ekonomi dalam satu indikator komprehensif (Ghislandi et al., 2018).

Menurut United Nations Development Programme (UNDP), IPM dihitung berdasarkan tiga dimensi utama, yaitu kesehatan yang diukur melalui Usia Harapan Hidup saat Lahir (UHH), pendidikan yang diwakili oleh Harapan Lama Sekolah (HLS) dan Rata-rata Lama Sekolah (RLS), serta standar hidup melalui Pengeluaran per Kapita yang Disesuaikan (PPP). Indikator-indikator ini memberikan gambaran komprehensif mengenai kualitas hidup dan kemajuan sosial suatu wilayah.

Provinsi Jawa Tengah sebagai salah satu wilayah dengan jumlah kabupaten/kota terbanyak di Indonesia memerlukan analisis terhadap pola IPM di Jawa Tengah untuk memetakan kesenjangan dan merumuskan kebijakan yang tepat sasaran. Namun, pendekatan konvensional dalam analisis kluster sering kali mengabaikan kompleksitas struktur data multidimensi dari indikator IPM.

Beberapa penelitian terdahulu telah mencoba mengatasi masalah ini. Studi oleh Sari et al. (2022) mengelompokkan kabupaten/kota di Provinsi Jawa Tengah berdasarkan komponen pembentuk IPM menggunakan algoritma K-Means dan metode *Elbow* untuk menentukan jumlah kluster optimal. Namun, penelitian ini belum memanfaatkan teknik reduksi dimensi sehingga berpotensi kehilangan informasi struktural data yang kompleks.

Di sisi lain, Mutiah et al. (2024) membandingkan kinerja K-Means dan DBSCAN dalam mengidentifikasi kelompok rumah tangga berdasarkan fasilitas sosial-ekonomi di Jawa Barat. Hasilnya menunjukkan bahwa K-Means lebih stabil untuk data berdistribusi globular, sedangkan DBSCAN lebih baik dalam mendeteksi *outlier*. Namun, kedua metode tersebut tidak menerapkan dengan reduksi dimensi sehingga interpretasi hasil klusterisasi menjadi terbatas.

Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan pendekatan inovatif dengan menggabungkan *Uniform Manifold Approximation and Projection* (UMAP) dan K-Means. UMAP merupakan teknik reduksi dimensi yang dapat menangkap hubungan nonlinier kompleks dan mempertahankan struktur lokal serta global dalam data (Mittal et al., 2024). Penggunaan kombinasi UMAP dan K-Means dalam konteks analisis IPM di tingkat kabupaten/kota di Jawa Tengah belum banyak dilakukan sehingga penelitian ini mengisi gap pada studi sebelumnya yang umumnya menggunakan metode klusterisasi tradisional tanpa reduksi dimensi yang memadai.

Dengan pendekatan ini, penelitian diharapkan dapat memberikan kontribusi signifikan baik dalam pengembangan metode analisis data pembangunan manusia maupun dalam perumusan kebijakan pembangunan yang lebih tepat sasaran di Provinsi Jawa Tengah.

KAJIAN TEORI

ELBOW METHOD

Elbow method merupakan metode yang digunakan untuk menentukan jumlah cluster terbaik, yaitu dengan cara melihat persentase setiap kluster yang akan membentuk siku pada titik tertentu (Maori & Evanita, 2023). Tujuan dari metode *elbow* adalah untuk memilih nilai k yang kecil dan memiliki nilai *withinss* (SSE) yang rendah (Maori & Evanita, 2023). Metode *elbow* biasanya disajikan dalam bentuk diagram untuk memudahkan penentuan jumlah cluster. Lokasi tikungan atau siku pada plot umumnya dianggap sebagai jumlah kluster yang tepat karena secara visual membentuk "siku" pada grafik (Kassambara, n.d.). Berikut adalah rumus untuk SSE (*Sum Square Error*) (Maori & Evanita, 2023):

$$SSE = \sum_{k=1}^k \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (1)$$

Keterangan:

C_k = K kluster yang terbentuk

k = banyak kluster

x_i = data x pada fitur ke- i

μ_k = rata-rata k kluster pada nilai k ($k = 1, 2, 3, \dots, k$)

SILHOUETTE SCORE

Nilai *silhouette* suatu objek menggambarkan seberapa baik objek tersebut cocok dengan klasternya sendiri (*cohesion*) dibandingkan dengan klaster lain (*separation*) (Tan et al., 2014). Penilaian dalam metode ini menggunakan metrik yang disebut *silhouette coefficient*, yang nilainya berkisar antara -1 sampai 1. Jika nilai *silhouette coefficient* mendekati 1, berarti objek tersebut sudah tergolong dengan tepat dalam klasternya, sedangkan nilai yang mendekati -1 menunjukkan bahwa objek tersebut lebih cocok berada di klaster lain (Saputra et al., 2020). Metode *silhouette* bertujuan untuk menemukan jumlah klaster (k) yang menghasilkan nilai *silhouette coefficient* tertinggi, yang biasanya dapat dilihat sebagai puncak tertinggi pada grafik. Berikut adalah rumus untuk *silhouette coefficient* (Rahannabil et al., 2024):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

Keterangan:

$a(i)$ = rata-rata jarak objek i ke semua objek dalam klaster sendiri

$b(i)$ = rata-rata objek i ke semua objek dalam klaster lain

STANDARISASI

Klasterisasi sangat bergantung pada ukuran dan skala variabel, terutama pada metode berbasis jarak seperti K-Means. Variabel dengan rentang nilai lebih besar akan memberikan pengaruh dominan dalam perhitungan jarak antar data sehingga dapat mendominasi hasil klasterisasi dan menyebabkan bias. Oleh karena itu, standarisasi memastikan bahwa setiap variabel memberikan kontribusi yang seimbang dalam proses analisis (Yapanto et al., 2023). Metode *Z-Score* mengubah nilai setiap atribut agar memiliki nilai rata-rata 0 dan deviasi standar 1 menggunakan rumus (Syasila & Fakhriza, 2025):

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

Keterangan:

x = nilai observasi

μ = rata-rata

σ = standar deviasi

Keunggulan *Z-Score* dibandingkan metode lain adalah tidak membatasi data pada rentang tertentu sehingga informasi variabilitas data tetap terjaga dengan baik. Dengan demikian, *Z-Score*

memungkinkan kontribusi variabel yang proporsional dalam proses klasterisasi sehingga hasil pengelompokan menjadi lebih akurat.

UNIFORM MANIFOLD APPROXIMATION AND PROJECTION (UMAP)

UMAP (*Uniform Manifold Approximation and Projection*) merupakan teknik reduksi dimensi nonlinear. Metode ini memproyeksikan data berdimensi tinggi ke dalam ruang berdimensi rendah dengan mempertahankan struktur topologi data asli. Berbeda dengan *Principal Component Analysis* (PCA) yang hanya efektif untuk data linear, UMAP mampu menangani struktur data nonlinear melalui pendekatan geometri manifold.

UMAP mengadopsi konsep dari topologi aljabar dan teori kategori. Metode ini terdiri dari dua tahap utama, yakni *Fuzzy Simplicial Set* dan Optimisasi *Cross-Entropy* (McInnes et al., 2020).

KONSTRUKSI FUZZY SIMPLICIAL SET

Pada ruang dimensi tinggi, hubungan antar titik dimodelkan sebagai kompleks simplicial kabur (*fuzzy simplicial set*) dengan fungsi kekaburan (McInnes et al., 2020):

$$v_{j|i} = \exp\left(-\frac{d(x_i x_j) - \rho_i}{\sigma_i}\right) \quad (4)$$

Keterangan:

$v_{j|i}$ = Bobot keterhubungan antara titik

ρ_i = Jarak ke tetangga terdekat pertama

σ_i = Parameter penskalaan lokal

OPTIMISASI CROSS-ENTROPY

Proyeksi ke ruang dimensi rendah dilakukan dengan meminimalkan fungsi *cross-entropy* (McInnes et al., 2020):

$$C_{UMAP} = \sum_{i \neq j} v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) \log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \quad (5)$$

Keterangan:

v_{ij} = probabilitas keterhubungan dalam ruang dimensi tinggi

$w_{ij} = \left(1 + a \left\|y_i - y_j\right\|_2^{2b}\right)^{-1}$ = probabilitas dalam ruang dimensi rendah

a dan b = parameter yang dioptimalkan

K-MEANS

K-Means adalah algoritma klasterisasi yang bertujuan membagi data ke dalam (klaster)

berdasarkan kemiripan. Setiap data akan dikelompokkan ke kluster dengan *centroid* (titik pusat) terdekat, yang dihitung menggunakan jarak Euclidean (Mukhtar et al., 2024).

Semakin dekat jarak antar data, semakin besar kemungkinan mereka tergabung dalam kluster yang sama. Dengan demikian, K-Means berupaya mengelompokkan data sehingga setiap kluster terdiri dari data yang saling mirip. Berikut adalah rumus untuk jarak Euclidean (Siagian et al., 2025):

$$d(x, c) = \sqrt{\sum_{i=1}^n (x_i - c_i)^2} \quad (6)$$

Keterangan:

$d(x, c)$ = jarak antardata dan *centroid* kluster

x_i = nilai atribut dari data ke- i

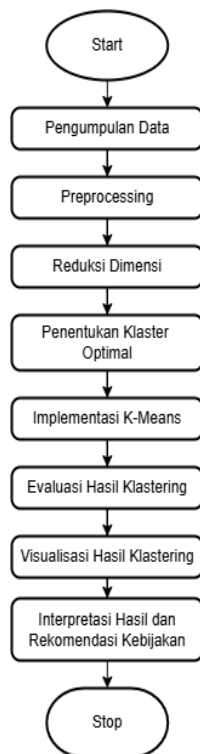
c_i = nilai *centroid* dari atribut ke- i

n = jumlah atribut dalam dataset

METODE

TAHAPAN PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan utama, dimulai dari pengumpulan data hingga interpretasi hasil dan rekomendasi kebijakan. Secara garis besar, alur penelitian dapat digambarkan sebagai berikut.



Gambar 1. Diagram Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama untuk mengidentifikasi segmentasi wilayah berdasarkan komponen Indeks Pembangunan Manusia (IPM). Tahap pertama adalah pengumpulan data, yakni dataset yang digunakan diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Tengah yang mencakup data dari 35 kabupaten/kota selama periode 2019–2024. Data ini mencakup variabel-variabel penting, seperti Angka Harapan Hidup, Rata-rata Lama Sekolah, Harapan Lama Sekolah, dan Pengeluaran per Kapita, yang digunakan untuk menganalisis pembangunan manusia. Setelah data terkumpul, tahapan berikutnya adalah *preprocessing* data, yang mencakup deteksi dan penanganan *missing values* dan *outliers*. Proses ini juga melibatkan standarisasi data dengan menggunakan metode *Z-Score* agar setiap fitur memiliki skala yang seragam dan relevan untuk analisis klusterisasi.

Setelah tahap *preprocessing*, dilakukan reduksi dimensi dengan metode *Uniform Manifold Approximation and Projection* (UMAP). Metode ini dipilih karena kemampuannya untuk mempertahankan struktur data yang kompleks dan nonlinier, yang lebih efektif daripada teknik linier seperti *Principal Component Analysis* (PCA). UMAP menghasilkan proyeksi dua dimensi yang digunakan untuk klusterisasi dan visualisasi data lebih lanjut. Penentuan jumlah kluster optimal menjadi langkah berikutnya, yang dilakukan dengan menggunakan dua metode utama, yaitu *Elbow Method* dan *Silhouette Score*. Kedua metode ini digunakan untuk memastikan kluster yang terbentuk memiliki kualitas yang baik dan representatif.

Selanjutnya, algoritma K-Means digunakan untuk mengelompokkan kabupaten/kota berdasarkan indikator pembangunan manusia. Proses klusterisasi dilakukan dengan memilih jumlah kluster yang optimal berdasarkan analisis sebelumnya. Setelah klusterisasi, hasilnya dievaluasi menggunakan *Silhouette Score* untuk mengukur kesesuaian objek dalam kluster serta pemisahan antarkluster. Hasil klusterisasi kemudian divisualisasikan menggunakan teknik UMAP, yang memberikan representasi dua dimensi dari distribusi kabupaten/kota di Jawa Tengah. Visualisasi ini memungkinkan pemahaman yang

lebih jelas tentang pola distribusi dan segmentasi wilayah.

Tahap akhir penelitian ini adalah interpretasi hasil klasterisasi untuk memahami karakteristik pembangunan manusia di Jawa Tengah, serta penyusunan rekomendasi kebijakan guna mendorong pemerataan dan peningkatan kualitas hidup di tiap wilayah. Penelitian ini juga membuka peluang komparasi dengan data IPM tahun berikutnya, termasuk 2025 dan seterusnya, untuk mengevaluasi perkembangan pola pembangunan secara longitudinal. Dengan demikian, hasil ini diharapkan menjadi dasar kuat bagi kebijakan pembangunan wilayah yang adaptif dan berbasis data jangka panjang.

PENGUMPULAN DATA

Penelitian ini menggunakan data sekunder dari Badan Pusat Statistik (BPS) Provinsi Jawa Tengah, meliputi 35 kabupaten/kota pada periode 2019–2024. Data terdiri dari empat variabel utama Indeks Pembangunan Manusia: Angka Harapan Hidup, Rata-rata Lama Sekolah, Harapan Lama Sekolah, dan Pengeluaran per Kapita. Variabel-variabel tersebut digabungkan dalam satu dataset untuk analisis longitudinal guna mengamati pola perkembangan tiap daerah dari tahun ke tahun. Sumber data berasal dari publikasi resmi BPS Jawa Tengah (Badan Pusat Statistik Provinsi Jawa Tengah, 2025). Hasil klasterisasi tidak hanya relevan untuk membandingkan dengan data 2025, tetapi juga menjadi acuan evaluasi dan pemantauan pembangunan manusia di masa depan.

PREPROCESSING

1. Deteksi dan Penanganan *Missing Values*

Langkah awal dalam *preprocessing* adalah memastikan kelengkapan data. Seluruh variabel numerik diperiksa untuk mengidentifikasi adanya nilai yang hilang (*missing values*). Jika ditemukan, penanganan dilakukan menggunakan metode *mean imputation*, yaitu dengan menggantikan nilai kosong dengan rata-rata dari variabel terkait. Pendekatan ini dipilih karena mampu mempertahankan distribusi data tanpa mengubah skala secara signifikan.

2. Deteksi dan Penanganan *Outlier*

Outlier terdeteksi melalui metode visualisasi *box plot* dan perhitungan *Z-Score* untuk mengidentifikasi nilai yang berada jauh dari distribusi normal.

Namun, karena data mencerminkan karakteristik unik masing-masing kabupaten/kota, *outlier* tidak dihapus guna menjaga keutuhan informasi spasial dan temporal.

3. Standarisasi Data

Data distandarisasi menggunakan metode *Z-Score*, yaitu dengan mengubah nilai setiap variabel menjadi distribusi dengan rata-rata nol dan standar deviasi satu. Proses normalisasi ini bertujuan untuk memastikan bahwa setiap fitur memiliki kontribusi yang seimbang dalam proses klasterisasi, serta mencegah dominasi dari variabel yang memiliki rentang nilai lebih besar.

4. Reduksi Dimensi

Reduksi dimensi dilakukan menggunakan metode *Uniform Manifold Approximation and Projection* (UMAP) karena hasil eksplorasi awal melalui *heatmap* korelasi dan *pairplot* menunjukkan bahwa hubungan antarvariabel cenderung bersifat nonlinier. Metode linier, seperti *Principal Component Analysis* (PCA) dianggap kurang mampu merepresentasikan pola tersembunyi dalam data sehingga UMAP dipilih karena lebih unggul dalam mempertahankan struktur lokal dan global. Proyeksi dua dimensi dari UMAP ini kemudian digunakan dalam tahap klasterisasi dan visualisasi, untuk memetakan distribusi wilayah berdasarkan indikator pembangunan manusia secara lebih efektif.

5. Penentuan Jumlah Klaster Optimal

Sebelum algoritma K-Means diterapkan, penentuan jumlah klaster yang optimal menjadi aspek penting agar proses klasterisasi mampu merepresentasikan segmentasi wilayah secara akurat. Dalam penelitian ini, jumlah klaster ditentukan menggunakan dua metode utama, yaitu *Elbow Method* dan *Silhouette Score*. Kedua metode ini digunakan untuk memastikan klaster yang terbentuk memiliki kualitas yang baik dan representatif.

a. *Elbow Method*

Metode ini mengidentifikasi jumlah klaster optimal dengan mengamati penurunan *inertia* (total jarak antara setiap titik data dengan *centroid* klaster). Titik di mana penurunan *inertia* mulai melambat (membentuk "siku" atau *elbow* pada grafik) dianggap sebagai jumlah klaster terbaik karena penambahan klaster setelah titik ini

tidak memberikan perbaikan signifikan pada model klasterisasi. Metode ini membantu menghindari *overfitting* dengan memilih jumlah klaster yang efisien dan bermakna secara statistik.

b. *Silhouette Score*

Silhouette Score mengukur seberapa baik sebuah titik data cocok dengan klasternya sendiri dibandingkan dengan klaster lain. Nilai skor yang tinggi menunjukkan bahwa data dalam satu klaster memiliki kemiripan yang tinggi dan terpisah dengan baik dari klaster lain. Skor ini dihitung berdasarkan jarak rata-rata antartitik dalam klaster dan jarak rata-rata ke titik di klaster terdekat lainnya sehingga memberikan gambaran kualitas klaster secara keseluruhan. Penggunaan *Silhouette Score* membantu dalam memilih jumlah klaster yang menghasilkan segmentasi data paling representatif.

6. Implementasi K-Means

Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan kabupaten/kota berdasarkan indikator pembangunan manusia. K-Means adalah metode *unsupervised learning* yang membagi data ke dalam klaster-klaster berdasarkan kemiripan fitur, dengan tujuan mengidentifikasi pola dalam data. Proses klasterisasi dimulai dengan pemilihan jumlah klaster (k) dan *centroid* awal secara acak. Setiap data dialokasikan ke klaster terdekat menggunakan *Euclidean Distance*, kemudian *centroid* diperbarui berdasarkan rata-rata posisi seluruh anggota klaster. Proses ini diulang hingga konvergen.

Dalam penelitian ini, jumlah klaster ditetapkan $k = 4$, berdasarkan analisis konteks data dan efektivitas visualisasi setelah reduksi dimensi. Hasil klasterisasi ini memberikan wawasan tentang pola kemiripan antarwilayah, yang digunakan sebagai dasar segmentasi pembangunan yang lebih terarah dan spesifik.

7. Evaluasi Hasil Klasterisasi

Evaluasi klaster menggunakan *Silhouette Score*, yang menilai kesesuaian objek dalam klaster dibandingkan dengan klaster lain. Skor mendekati 1 menunjukkan pemisahan klaster yang baik. Metrik ini digunakan untuk mengukur keberhasilan

pengelompokan wilayah dalam membedakan karakteristik antar kabupaten/kota.

8. Visualisasi Hasil Klasterisasi

Visualisasi hasil klasterisasi dilakukan melalui proyeksi dua dimensi menggunakan UMAP, yang merepresentasikan 35 kabupaten/kota di Jawa Tengah berdasarkan indikator pembangunan manusia. Klaster hasil algoritma K-Means ditampilkan dalam bentuk *scatter plot* untuk mengungkap pola kemiripan antardaerah secara visual.

Tujuan utama visualisasi ini adalah untuk memetakan struktur klaster secara jelas, mengidentifikasi pola spasial yang konsisten dari tahun ke tahun (2019–2024), serta menyediakan dasar komparatif terhadap data IPM tahun 2025. Dengan demikian, analisis ini tidak hanya mendeskripsikan segmentasi wilayah, tetapi juga mendukung perumusan kebijakan pembangunan yang lebih berbasis data.

HASIL DAN PEMBAHASAN

PENGUMPULAN DATA

Dalam penelitian ini, digunakan data yang bersumber dari Badan Pusat Statistik (BPS) dengan judul “Indeks Pembangunan Manusia Menurut Kabupaten/Kota, 2019–2024”. Data tersebut disusun berdasarkan standar format BPS dan mencakup berbagai variabel yang tidak seluruhnya relevan dengan tujuan penelitian. Oleh karena itu, dilakukan proses penyesuaian variabel dan struktur data agar sesuai dengan kebutuhan analisis. Gambar berikut menyajikan ringkasan data yang telah disesuaikan.

Kabupaten/Kota	Tahun	x_1	x_2	x_3	x_4
Cilacap	2019	73,52	12,49	6,93	10639
Banyumas	2019	73,55	12,82	7,42	11703
Purbalingga	2019	73,02	11,98	7,14	10131
Banjarnegara	2019	74,01	11,45	6,50	9547
Kebumen	2019	73,22	13,04	7,53	9066
...
Tegal	2024	74,87	13,25	9,28	14358

Gambar 2. Tampilan Dataset

Keterangan:

x_1 = Usia Harapan Hidup saat Lahir

x_2 = Harapan Lama Sekolah

x_3 = Rata-Rata Lama Sekolah

x_4 = Pengeluaran per Kapita

Data tersebut diperoleh dari penghapusan variabel yang tidak relevan serta penggabungan

data dari tahun 2019 hingga 2024 untuk membentuk satu kesatuan data yang siap dianalisis.

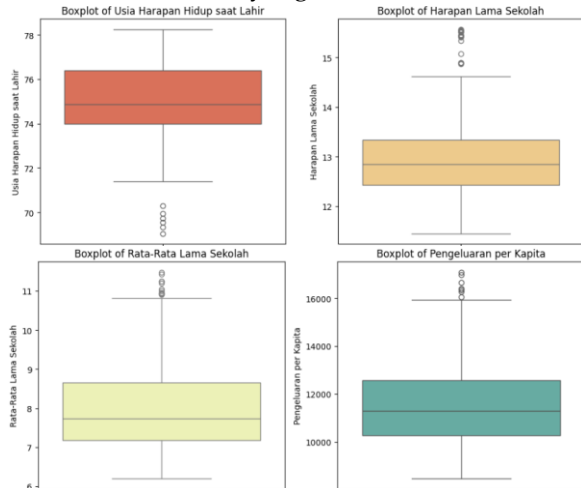
PREPROCESSING

1. Deteksi dan Penanganan Missing Values

Hasil pemeriksaan menunjukkan bahwa 210 entri dalam dataset tercatat secara lengkap sehingga tidak terdapat nilai yang hilang atau *missing values*. Kondisi ini memastikan bahwa data siap untuk dianalisis tanpa memerlukan proses imputasi atau penanganan data yang tidak lengkap.

2. Deteksi dan Penanganan Outlier

Deteksi *outlier* dengan metode *Z-Score* menunjukkan adanya nilai yang menyimpang dari kumpulan data utama. *Box plot* berikut menyajikan visualisasi dari *outlier* yang terdeteksi.



Gambar 3. Visualisasi *Box plot* untuk Mendeteksi *Outlier*

Box plot di atas mengindikasikan adanya nilai *outlier* pada keempat variabel yang digunakan. Namun, dalam penelitian ini setiap data dinilai memiliki informasi penting yang mencerminkan karakteristik dari masing-masing kabupaten/kota. Oleh karena itu, nilai *outlier* pada dataset tidak akan dihapus atau diolah lebih lanjut agar nilai asli dapat dipertahankan.

3. Standarisasi Data

Proses standarisasi dilakukan dengan menggunakan nilai *Z-Score* untuk mengubah data sehingga memiliki rata-rata 0 dan standar deviasi 1. Tabel berikut menyajikan hasil standarisasi dari masing-masing variabel.

Tabel 1. Data setelah Standarisasi

x_1	x_2	x_3	x_4
-0,8575	-0,5365	-0,8904	-0,5033
-0,8410	-0,1764	-0,5031	0,0744

-1,1318	-1,0931	-0,7244	-0,7791
-0,5887	-1,6715	-1,2302	-1,0962
-1,0220	0,0636	-0,4162	-1,3574
...
-0,1169	0,2927	0,9668	1,5161

Keterangan:

x_1 = Usia Harapan Hidup saat Lahir

x_2 = Harapan Lama Sekolah

x_3 = Rata-Rata Lama Sekolah

x_4 = Pengeluaran per Kapita

Hasil standarisasi menunjukkan bahwa data yang sebelumnya berada dalam rentang yang sangat bervariasi kini memiliki sebaran nilai yang lebih seragam. Kondisi ini memudahkan proses analisis lanjutan, seperti analisis kluster dengan menerapkan algoritma K-Means.

4. Reduksi Dimensi

Reduksi dimensi menggunakan metode UMAP (*Uniform Manifold Approximation and Projection*) dilakukan untuk meringkas empat variabel awal menjadi dua dimensi utama pada data yang bersifat nonlinier. Hasil dari proses reduksi dimensi disajikan pada tabel berikut.

Tabel 2. Data setelah Reduksi Dimensi

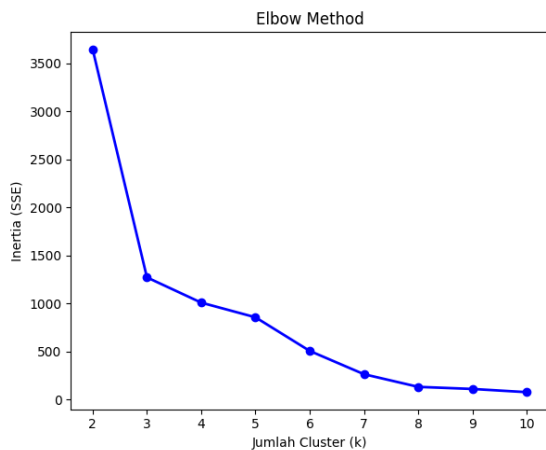
UMAP 1	UMAP 2
-0,8698	6,3777
-1,9479	5,6459
-0,4747	6,9827
-5,1063	8,4718
2,2809	2,1956
...	...
-3,9571	2,7000

Hasil tersebut diperoleh dengan memanfaatkan parameter terbaik UMAP untuk dataset ini, yaitu $n_neighbors = 10$ dan $min_dist = 0,0001$ yang dihitung berdasarkan jarak Euclidean. Parameter tersebut dirancang untuk mempertahankan struktur lokal data berdasarkan sepuluh tetangga terdekat sekaligus memungkinkan penyusunan titik-titik dalam kluster dengan jarak seminimal mungkin.

5. Penentuan Jumlah Cluster

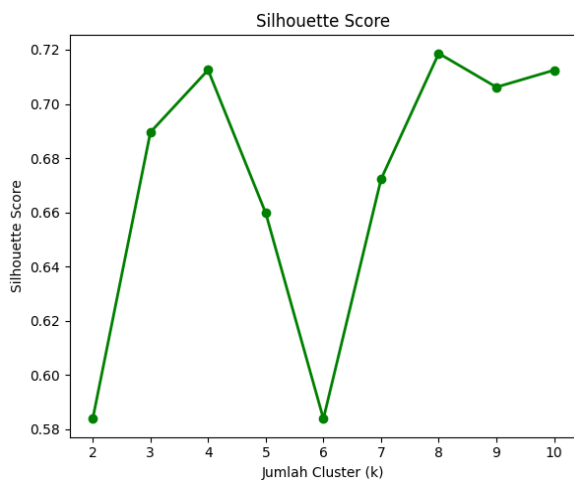
Jumlah kluster ditentukan berdasarkan analisis penurunan nilai *inertia* yang divisualisasikan melalui

Elbow Method, yang membantu mengidentifikasi titik optimal di mana penambahan jumlah kluster tidak lagi memberikan penurunan inertia yang signifikan. Untuk memperkuat keputusan tersebut, evaluasi tambahan dilakukan menggunakan *Silhouette Score*, yang mengukur seberapa baik setiap objek berada dalam klasternya masing-masing. Visualisasi dari *Elbow Method* dan *Silhouette Score* disajikan pada gambar berikut sebagai dasar pemilihan jumlah kluster yang optimal.



Gambar 4. Visualisasi *Elbow Method*

Berdasarkan visualisasi di atas, dapat dilihat bahwa nilai *inertia* mengalami penurunan signifikan hingga kluster berjumlah tiga. Setelah itu, grafik perlahan melandai dan membentuk lekukan yang menyerupai siku. Hal ini mengindikasikan bahwa jumlah kluster yang direkomendasikan untuk dataset ini adalah tiga. Di sisi lain, pengecekan dengan nilai *Silhouette Score* menunjukkan nilai yang sedikit berbeda. Visualisasi dari *Silhouette Score* akan ditampilkan pada gambar berikut.



Gambar 5. Visualisasi *Silhouette Score*

Visualisasi ini menunjukkan bahwa nilai *Silhouette Score* tertinggi tercapai saat kluster berjumlah empat dan delapan, dengan *score* lebih dari 0,70. Nilai yang cukup tinggi ini mengindikasikan bahwa pemisahan antarkluster berlangsung dengan baik. Dengan mempertimbangkan dua metode penentuan kluster di atas, diperoleh kesimpulan bahwa jumlah kluster yang paling optimal dalam analisis ini adalah empat.

IMPLEMENTASI K-MEANS

K-Means diimplementasikan dengan jumlah kluster yang paling optimal, yaitu $k = 4$. Tabel di bawah ini menyajikan jumlah anggota setiap kluster dari perhitungan yang telah dilakukan.

Tabel 3. Jumlah Anggota Kluster

Kluster	Jumlah
0	60
1	102
2	30
3	18

Hasil klusterisasi menunjukkan bahwa distribusi jumlah anggota pada setiap kluster cukup bervariasi. Variasi ini mencerminkan adanya perbedaan karakteristik dan pola data yang signifikan antarkluster. Dengan demikian, analisis lebih lanjut akan difokuskan pada pemahaman dan interpretasi karakteristik unik dari setiap kluster yang telah terbentuk.

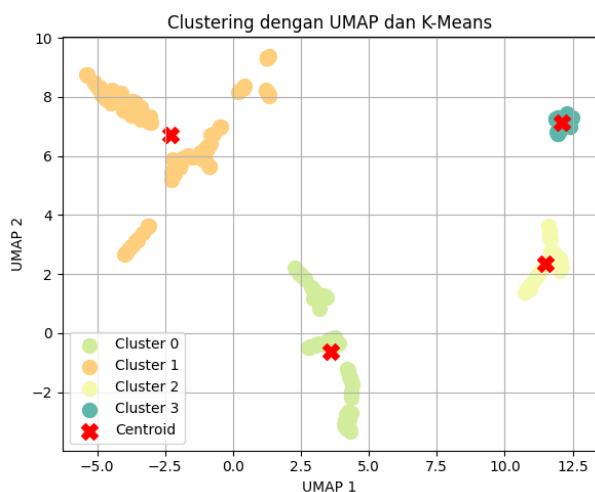
EVALUASI HASIL KLATERISASI

Silhouette Score menunjukkan nilai sebesar 0,7125. Nilai ini mengindikasi bahwa sebagian besar kabupaten/kota telah dikelompokkan ke dalam kluster yang sesuai dengan karakteristiknya. Dengan kata lain, sesama anggota dalam suatu kluster memiliki kesamaan yang identik sementara anggota antar kluster memiliki perbedaan yang signifikan. Oleh karena itu, dapat dibuktikan bahwa proses klusterisasi telah berjalan secara efektif dalam memetakan kabupaten/kota di Jawa Tengah berdasarkan pola Indeks Pembangunan Manusia (IPM).

VISUALISASI HASIL KLATERISASI

Selain perhitungan kuantitatif melalui *Silhouette Score*, visualisasi hasil klusterisasi juga penting untuk menilai kualitas pengelompokan secara visual.

Melalui visualisasi dua dimensi, dapat diamati sejauh mana objek-objek dengan karakteristik serupa dikelompokkan bersama, serta sejauh mana objek yang berbeda berhasil dipisahkan antar kluster. Visualisasi berikut memberikan gambaran intuitif mengenai struktur dan pola dari hasil klasterisasi yang diperoleh.



Gambar 6. Visualisasi Hasil Klasterisasi

Visualisasi di atas menggambarkan bahwa setiap kluster dipisahkan dengan sangat baik. Hal ini dapat dilihat melalui jarak yang terpaut cukup jauh antara satu kluster dengan kluster lain serta jarak yang sangat dekat antara anggota dalam satu kluster. Di samping itu, dapat diamati pula kepadatan dari masing-masing kluster. Kluster 3 tampak paling padat dan terfokus pada satu bagian sehingga mencerminkan homogenitas varians di dalamnya. Sementara itu, tiga kluster lain tampak lebih tersebar, yang mengindikasikan adanya variasi internal.

INTERPRETASI HASIL

Hasil klasterisasi menunjukkan bahwa kabupaten/kota di provinsi Jawa Tengah dapat dikelompokkan ke dalam empat kluster berdasarkan indikator Indeks Pembangunan Manusia (IPM). Untuk memahami karakteristik masing-masing kluster, dapat dilakukan analisis rata-rata (*mean*) dari setiap variabel. Tabel berikut menyajikan ringkasan nilai rata-rata sebagai representasi dari masing-masing kluster.

Tabel 4. Mean Setiap Kluster

Kluster	x_1	x_2	x_3	x_4
0	75,623	12,906	7,789	11110,15

1	73,693	12,416	7,327	10906,16
2	77,223	13,681	9,404	12140,06
3	77,590	15,271	10,830	15866,83

Keterangan:

x_1 = Usia Harapan Hidup saat Lahir

x_2 = Harapan Lama Sekolah

x_3 = Rata-Rata Lama Sekolah

x_4 = Pengeluaran per Kapita

Berdasarkan tabel di atas, dapat diketahui bahwa kluster 3 merupakan kelompok kabupaten/kota dengan indikator pembangunan manusia yang paling tinggi di Provinsi Jawa Tengah. Kluster ini memiliki rata-rata usia harapan hidup saat lahir sebesar 77,59 tahun, harapan lama sekolah 15,27 tahun, rata-rata lama sekolah 10,83 tahun, dan pengeluaran per kapita mencapai Rp15.866,8. Rata-rata tersebut merupakan angka tertinggi dibandingkan kluster lain. Capaian ini mengindikasikan bahwa kluster 3 merepresentasikan wilayah dengan kualitas hidup yang sangat baik, baik dari sisi kesehatan, pendidikan, maupun kesejahteraan ekonomi. Umumnya, kabupaten/kota dalam kluster 3 merupakan kota besar yang menjadi pusat perekonomian atau pemerintahan sehingga memiliki akses terhadap fasilitas pendidikan, kesehatan, serta kesempatan ekonomi yang lebih merata dan berkembang.

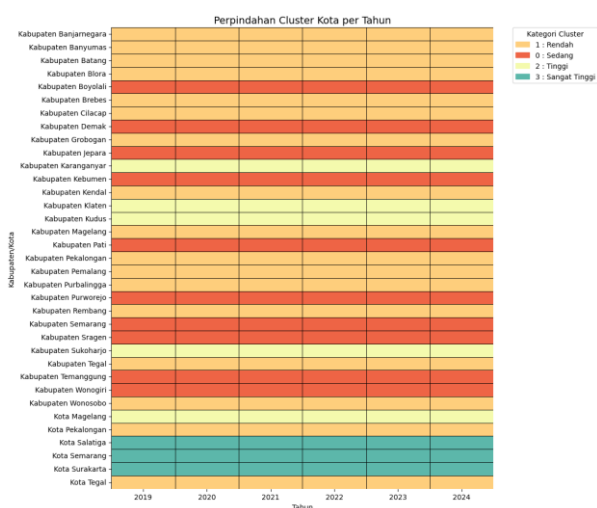
Sementara itu, kluster 2 mencerminkan kabupaten/kota dengan tingkat pembangunan manusia yang tergolong tinggi. Nilai rata-rata indikatornya mencapai 77,22 tahun untuk usia harapan hidup saat lahir, 13,68 tahun dan 9,40 tahun untuk harapan lama sekolah dan rata-rata lama sekolah, serta Rp12.140,06 untuk pengeluaran per kapita. Meski tidak setinggi kluster 3, nilai rata-rata tersebut menunjukkan bahwa wilayah dalam kluster 2 telah memiliki pondasi pembangunan yang cukup kuat. Artinya, kabupaten/kota yang tergabung pada kluster ini mungkin berlokasi tidak terlalu jauh dari pusat pertumbuhan ekonomi sehingga tetap memperoleh dampak positif dari pembangunan yang lebih terdistribusi.

Kluster 0 menggambarkan wilayah dengan pembangunan manusia yang tergolong menengah. Kluster ini memiliki rata-rata usia harapan hidup sebesar 75,62 tahun, harapan lama sekolah sebesar

12,91 tahun, rata-rata lama sekolah sebesar 7,79 tahun, dan pengeluaran per kapita sebesar Rp11.110,15. Angka ini menunjukkan bahwa pencapaian kabupaten/kota dalam klaster 0 berada pada tingkat yang cukup, meskipun belum optimal. Artinya, klaster tersebut mencerminkan daerah-daerah yang mungkin berada di kawasan penyangga atau wilayah transisi antara pusat dan pinggiran. Dengan demikian, penyediaan akses layanan dasar pada kabupaten/kota dalam klaster 0 memerlukan dorongan yang lebih kuat agar pembangunan manusia dapat berlangsung secara merata.

Di sisi lain, klaster 1 merupakan kelompok kabupaten/kota dengan capaian indikator pembangunan manusia yang paling rendah di antara klaster lain. Rata-rata usia harapan hidup hanya menunjukkan angka 73,69 tahun, harapan lama sekolah sebanyak 12,42 tahun, rata-rata lama sekolah sebanyak 7,33 tahun, dan pengeluaran per kapita sebesar Rp10.906,16. Kondisi ini mengindikasikan adanya tantangan besar dalam pemenuhan kebutuhan dasar masyarakat, baik dalam bidang kesehatan, pendidikan, maupun kesejahteraan ekonomi. Wilayah yang tergabung dalam klaster ini umumnya terletak di daerah pinggiran, jauh dari pusat pertumbuhan, dan masih didominasi oleh karakteristik pedesaan dengan keterbatasan akses terhadap teknologi dan layanan publik yang memadai.

Guna memberikan gambaran yang lebih jelas mengenai persebaran klaster serta perubahan posisi kabupaten/kota selama periode 2019 hingga 2024, disajikan visualisasi berikut.



Gambar 7. Perpindahan Klaster per Tahun

Dari visualisasi diatas, dapat diamati bahwa 17 kabupaten/kota di provinsi Jawa Tengah termasuk ke dalam klaster 1, yaitu wilayah dengan tingkat pembangunan manusia yang tergolong rendah. Jumlah tersebut terdiri atas 15 kabupaten yang meliputi Banjarnegara, Banyumas, Batang, Blora, Brebes, Cilacap, Grobogan, Kendal, Magelang, Pekalongan, Pemalang, Purbalingga, Rembang, Tegal, dan Wonosobo, serta dua wilayah perkotaan yakni Kota Pekalongan dan Kota Tegal. Kondisi ini menunjukkan bahwa sebagian besar wilayah di Jawa Tengah masih menghadapi tantangan dalam pemenuhan indikator utama pembangunan manusia, seperti akses pendidikan, layanan kesehatan, dan kesejahteraan ekonomi. Oleh karena itu, diperlukan pembentukan kebijakan yang lebih intensif dan terfokus untuk meningkatkan kualitas layanan dasar dan kapasitas ekonomi secara lokal.

Sementara itu, hanya terdapat tiga wilayah di Provinsi Jawa Tengah dengan tingkat pembangunan manusia yang sangat tinggi, yakni Kota Salatiga, Kota Semarang, dan Kota Surakarta. Ketiga kota ini merupakan wilayah yang memiliki peran strategis dalam pertumbuhan regional. Kondisi ini mencerminkan adanya ketimpangan antarwilayah, di mana wilayah perkotaan cenderung lebih unggul dalam pencapaian indikator indeks pembangunan manusia dibandingkan daerah penyangga atau pinggiran. Pemerintah daerah dapat menjadikan ketiga kota ini sebagai teladan dalam merencanakan kebijakan pembangunan bagi wilayahnya masing-masing. Dengan demikian, pemerataan terhadap seluruh indikator indeks pembangunan manusia dapat terlaksana dengan baik.

Di sisi lain, visualisasi di atas juga memperlihatkan tidak adanya perpindahan klaster bagi setiap kabupaten/kota di Provinsi Jawa Tengah. Hal ini mengindikasikan bahwa setiap wilayah mampu mempertahankan kualitasnya pada seluruh indikator Indeks Pembangunan Manusia (IPM). Kondisi tersebut menjadi kabar baik bagi wilayah dengan tingkat pembangunan manusia yang tergolong tinggi. Sebaliknya, wilayah yang tergolong dalam klaster rendah dan sedang justru menunjukkan tidak adanya perkembangan pembangunan dalam aspek pendidikan, kesehatan, dan ekonomi. Pemerintah pusat maupun daerah perlu mengevaluasi kebijakan pembangunan pada ketiga aspek tersebut sehingga pembangunan di

Provinsi Jawa Tengah dapat berjalan secara efektif, merata, dan berkelanjutan. Upaya kolaboratif dan inovatif sangat dibutuhkan untuk mengatasi tantangan tersebut dan mendorong percepatan pembangunan di seluruh wilayah.

PENUTUP

SIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma K-Means dalam pengelompokan 35 kabupaten/kota di Provinsi Jawa Tengah berdasarkan empat indikator utama Indeks Pembangunan Manusia (IPM), yaitu Usia Harapan Hidup, Harapan Lama Sekolah, Rata-rata Lama Sekolah, dan Pengeluaran per Kapita, untuk periode 2019-2024. Melalui tahap *preprocessing*, standarisasi, dan reduksi dimensi menggunakan UMAP, ditemukan bahwa pembentukan empat kluster merupakan konfigurasi yang optimal dengan nilai *Silhouette Score* tertinggi. Hasil klusterisasi menunjukkan adanya perbedaan signifikan dalam tingkat pembangunan manusia antardaerah, di mana Kluster 3 merepresentasikan wilayah dengan kualitas pembangunan tertinggi, sedangkan Kluster 0 mencerminkan daerah dengan tantangan pembangunan yang lebih kompleks. Temuan ini memberikan gambaran nyata terhadap ketimpangan pembangunan di wilayah Jawa Tengah serta membuktikan bahwa analisis kluster dapat menjadi pendekatan yang efektif untuk mendukung perumusan kebijakan pembangunan yang berbasis data.

SARAN

Sebagai tindak lanjut dari penelitian ini, disarankan agar analisis klusterisasi diperbarui secara berkala dengan menyertakan data IPM terbaru, seperti tahun 2025 dan seterusnya, untuk memantau dinamika pembangunan di setiap kabupaten/kota secara lebih akurat. Penelitian lanjutan juga sebaiknya mempertimbangkan penambahan variabel relevan lainnya yang berkontribusi terhadap IPM, seperti akses terhadap layanan kesehatan, kualitas pendidikan, tingkat pengangguran, serta kondisi infrastruktur. Selain itu, eksplorasi terhadap metode klusterisasi alternatif seperti K-Means++, DBSCAN, atau algoritma berbasis hierarki dapat dilakukan guna memperoleh hasil kluster yang lebih stabil dan interpretatif.

Penggabungan pendekatan prediktif berbasis *machine learning* juga dapat dipertimbangkan untuk memberikan proyeksi pembangunan di masa mendatang sehingga hasil analisis tidak hanya bersifat deskriptif, tetapi juga bersifat strategis dalam perencanaan kebijakan pembangunan daerah yang berkelanjutan dan inklusif.

DAFTAR PUSTAKA

- Badan Pusat Statistik Provinsi Jawa Tengah. (2025). *[Metode baru] Indeks Pembangunan Manusia menurut kabupaten/kota*. <https://jateng.bps.go.id/id/statistics-table/2/ODMjMg==/-metode-baru--indeks-pembangunan-manusia-menurut-kabupaten-kota.html>
- Ghislandi, S., Sanderson, W. C., & Scherbov, S. (2018). A Simple Measure of Human Development: The Human Life Indicator. *Population and Development Review*, 45(1), 219-233. <https://doi.org/10.1111/padr.12205>
- Kassambara, A. (n.d.). Determining the optimal number of clusters: 3 must know methods. *Datanovia.com*. <https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>
- Maori, M. A., & Evanita. (2023). Metode Elbow dalam optimasi jumlah cluster pada K-Means clustering. *Jurnal Teknik Mesin, Elektro dan Ilmu Komputer (SIMETRIS)*, 14(2), 277-287. <https://doi.org/10.24176/simet.v14i2>
- McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. <https://arxiv.org/abs/1802.03426>
- Mittal, M., Gujjar, P. J., Prasad, G. M. S., Devadas, R. M., Ambreen, L., & Kumar, V. (2024). Dimensionality reduction using UMAP and t-SNE technique. *Proceedings of the 2024 Second International Conference on Advances in Information Technology (ICAIT)*, 1-5. <https://doi.org/10.1109/ICAIT61638.2024.10690797>
- Mukhtar, H., Pramaditya, I. D., Weisdiyanto, W. S., Putra, S. H., Trimuawasih, D., & Rilda, A. A. (2024). Algoritma K-Means untuk pengelompokan perilaku customer. *Jurnal Software Engineering and Information System (SEIS)*, 4(2), 96-101. <https://doi.org/10.37859/seis.v4i2.7615>
- Mutiah, S., Hasnataeni, Y., Fitrianto, A., Erfiani, & Jurmansyah, L. M. R. D. (2024). Perbandingan Metode Klustering K-Means dan DBSCAN dalam Identifikasi Kelompok Rumah Tangga

Berdasarkan Fasilitas Sosial Ekonomi di Jawa Barat. *Teorema: Teori dan Riset Matematika*, 9(2), 247-260.

- Rahannabil, S. D., Ilyas, H. M. A., Shafira, H. N., Riam, M. A., Hastim, N. N., & Siregar, T. K. H. (2024). Perbandingan Agglomerative Nesting dan K-Means untuk klasterisasi ketimpangan gender berdasarkan dimensi kesehatan reproduksi. *Seminar Nasional Official Statistics 2024*, 459-470.
- Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. In *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*.
<https://doi.org/10.2991/aisr.k.200424.051>
- Sari, R. Y., Oktavianto, H., & Sulisty, H. W. (2022). Algoritma K-Means dengan metode Elbow untuk mengelompokkan kabupaten/kota di Jawa Tengah berdasarkan komponen pembentuk Indeks Pembangunan Manusia. *Jurnal Smart Teknologi*, 3(2), 104-108.
- Siagian, R. I. P., Pratama, E., Lubis, F. A., Priscilia, S. A., & Ramadhan, F. (2025). Segmentasi pelanggan dengan algoritma K-Means untuk strategi pemasaran yang efektif. *Jurnal Mahasiswa Teknik Informatika (JATI)*, 9(4), 5615-5620.
<https://doi.org/10.36040/jati.v9i4.13865>
- Syaqila, S., & Fakhri, M. (2025). K-Means clustering untuk mengukur pengaruh kompetensi terhadap kinerja pegawai. *Journal of Information System Research (JOSH)*, 6(2).
<https://doi.org/10.47065/josh.v6i2.6758>
- Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining*. Pearson.
- United Nations Development Programme (UNDP). (n.d.). Human Development Index (HDI). <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>
- Yapanto, L. M., Muzfirah, S., Aras, N. R. M., & Sibua, N. (2023). Analisis data statistik metode dan teknik. PT Media Penerbit Indonesia. <http://repository.mediapenerbitindonesia.com/id/eprint/301>