Volume 5 No. 1 (January - June 2025), pp. 32-48 | ISSN 2962-2360 (online) | ISSN 2962-3669 (print)
Arcitech: Journal of Computer Science and Artificial Intelligence
https://journal.iaincurup.ac.id/index.php/arcitech/article/view/13488 | DOI : https://dx.doi.org/10.29240/arcitech.v5i1.13488

Analisis Wilayah Prioritas Pembangunan di Provinsi Jawa Timur Berdasarkan Indikator Sosial Menggunakan Metode *K-Means Clustering*

Gita Rohma Utami Asyafiiyah[1], Artika Widyastuti[2], Friska Andriani[3]
[1,2,3]Program Studi Informatika, Universitas Islam Majapahit, Jawa Timur, Indonesia
Email : gitarohma7@gmail.com, artikawd005@gmail.com, friskaandriani7474@gmail.com

Article Information

Article history

Received 30 April 2025
Revised 20 May 2025
Accepted 20 June 2025
Available 30 June 2025

Keywords

Machine Learning
Clustering
K-Means
Development

Corresponding Author:

Gita Rohma Utami Asyafiiyah,
Universitas Islam Majapahit,
Email : gitarohma7@gmail.com

Abstract

Development is a key indicator of a region's progress; however, regional disparities remain a pressing issue due to uneven distribution of development. This study employs a data mining approach using the K-Means Clustering method. The objective is to classify priority development areas in East Java Province based on various social indicators, including total population, population growth rate, population density, Human Development Index (HDI), unemployment rate, and average years of schooling (AYS). Unlike previous studies, this research adopts a more comprehensive data-driven approach. The results show that K-Means successfully classifies regions into two clusters: priority and non-priority. A Silhouette Score of 0.45 indicates a fairly good level of cluster separation. Most of the regions in the priority cluster are regencies, while the non-priority cluster predominantly consists of urban areas. These findings confirm that K-Means Clustering is an effective decision-support tool for identifying priority development areas through data-based analysis.

Keywords : *Machine Learning, Clustering, K-Means, Development*

Abstrak

Pembangunan merupakan indikator utama kemajuan suatu daerah, namun ketimpangan antarwilayah masih menjadi persoalan akibat kurangnya pemerataan pembangunan. Penelitian ini menggunakan pendekatan data mining dengan metode K-Means Clustering. Penelitian ini bertujuan untuk mengelompokkan wilayah prioritas pembangunan di Provinsi Jawa Timur berdasarkan indikator sosial, seperti jumlah penduduk, laju pertumbuhan, kepadatan, IPM, tingkat pengangguran, dan RLS. Berbeda dari studi sebelumnya, penelitian ini mengadopsi pendekatan data yang lebih komprehensif. Hasilnya menunjukkan bahwa K-Means mampu membagi wilayah menjadi dua klaster: prioritas dan non-prioritas. Nilai Silhouette Score sebesar 0,45 mengindikasikan pemisahan klaster yang cukup baik. Mayoritas wilayah dalam klaster prioritas adalah daerah kabupaten, sementara klaster non-prioritas didominasi wilayah perkotaan. Temuan ini menegaskan bahwa K-Means Clustering efektif digunakan sebagai alat bantu pengambilan keputusan dalam menentukan wilayah prioritas pembangunan secara berbasis data.

Kata Kunci : *machine learning, clustering, k-means, pembangunan*

1. Pendahuluan

Pembangunan pada hakikatnya adalah proses perubahan yang berlangsung secara berkelanjutan ke arah yang lebih baik dengan berdasarkan pada norma-norma tertentu. Tujuan utama pembangunan ialah meningkatkan taraf hidup masyarakat secara menyeluruh, meliputi aspek pendidikan, kesehatan, kesejahteraan ekonomi, dan pemerataan sosial (Rinaldi, 2020). Dalam konteks pembangunan nasional, upaya pemerataan pembangunan menjadi salah satu fokus utama untuk menghindari terjadinya kesenjangan antar wilayah, khususnya di negara kepulauan seperti Indonesia yang memiliki karekteristik geografis dan sosial yang sangat beragam (Muhammad Fakhrur Rodzi, 2023).

Namun, pada kenyataannya ketimpangan sosial antar wilayah masih menjadi persoalan yang serius (Anwar et al., 2023). Beberapa daerah menunjukkan kemajuan pesat di bidang sosial dan ekonomi, sementara daerah lain tertinggal dengan berbagai permasalah sosial yang cukup serius, seperti tingginya angka kemiskinan, rendahnya tingkat pendidikan, serta terbatasnya akses terhadap berbagai layanan dasar. Hal ini tidak hanya berdampak pada penurunan kualitas hidup masyarakat, tetapi juga dapat menghambat tercapainya tujuan utama dari pembangunan berkelanjutan (Nurvia Handayani & Nurul Hanifa, 2024).

Kondisi ketimpangan ini tidak hanya terjadi antarprovinsi, tetapi juga dapat ditemukan di dalam satu provinsi yang memiliki jumlah wilayah administratif tinggi dengan karakteristik sosial yang beragam. Salah satunya adalah Provinsi Jawa Timur, yang merupakan provinsi terpadat kedua di Indonesia, yang terdiri dari 38 kabupaten dan kota (Amalia et al., 2024). Keberagaman sosial, ekonomi, dan demografi antar wilayah di provinsi ini menimbulkan tantangan tersendiri dalam mewujudkan pembangunan yang merata. Beberapa daerah di Jawa Timur menunjukkan capaian pembangunan yang cukup tinggi, sementara daerah lainnya masih menghadapi berbagai persoaalan sosial yang mendasar seperti kemiskinan, pengangguran, dan rendahnya akses pendidikan. Kondisi ini menunjukkan bahwa diperlukan adanya pendekatan analitis yang mampu memetakan wilayah secara objektif dan komprehesif berdasarkan indikator-indikator sosial yang relevan, agar kebijakan pembangunan dapat dilaksanakan dengan lebih terarah dan tepat sasaran.

Salah satu pendekatan potensial dalam mengatasi kebutuhan tersebut adalah penerapan teknik data mining. Meskipun teknik data mining seperti K-Means Clustering telah banyak diterapkan di berbagai sektor, penerapannya dalam konteks perencanaan pembangunan wilayah, khususnya untuk mengidentifikasi daerah prioritas berdasarkan indikator sosial, masih tergolong terbatas. Penelitian-penelitian terdahulu umumnya hanya menggunakan satu hingga tiga indikator dalam pemetaan wilayah, seperti kemiskinan, tingkat pengganggguran, atau Indeks Pembangunan Manusia (IPM), sehingga hasilnya belum mampu menggambarkan kompleksitas kondisi sosial secara

menyeluruh. Selain itu, sebagian besar penelitian sebelumnya juga hanya berfokus pada proses klasifikasi data atau validasi metode algoritmik, tanpa disertai tujuan yang jelas terkait pemanfaatan hasil identifikasi atau pemetaan wilayah yang dilakukan.

Berdasarkan latar belakang dan celah penelitian yang telah diuraikan sebelumnya, penelitian ini bertujuan untuk melakukan pemetaan wilayah prioritas pembangunan di Provinsi Jawa Timur dengan pendekatan berbasis data menggunakan metode K-Means Clustering. Pengelompokan wilayah dilakukan berdasarkan sejumlah indikator sosial yang lebih komprehesif, seperti jumlah penduduk, laju pertumbuhan penduduk, kepadatan penduduk, tingkat kemiskinan, tingkat pengangguran terbuka (TPT), indeks pembangunan manusia (IPM), dan rata-rata lama sekolah (RLS). Melalui pendekatan tersebut, penelitian ini diharapkan mampu menghasilkan klasifikasi wilayah yang objektif dan menyeluruh, tidak bergantung pada satu indikator tunggal, melainkan berdasarkan pola alami yang terbentuk dari data. Selain itu, penelitian ini juga berupaya mengungkap karakteristik sosial utama yang membedakan antar klaster wilayah, sebagai dasar untuk merumuskan arah kebijakan pembangunan yang lebih merata dan sesuai dengan kondisi masing-masing kelompok daerah. Dengan demikian, wilayah-wilayah yang memerlukan perhatian dan tindak lanjut kebijakan dapat diidentifikasi secara lebih akurat.

Penelitian ini juga diharapkan dapat memberikan kontribusi di bidang data science, khusunya dalam penerapan teknik data mining untuk analisis spasial dan sosial. Selain itu, hasil penelitian ini dapat menjadi referensi penting bagi pemerintah daerah dan pemangku kepentingan dalam menentukan wilayah yang membutuhkan perhatian pembangunan lebih awal. Secara keseluruhan, penelitian ini tidak hanya menjawab kebutuhan akan pemetaan wilayah berbasis data, tetapi juga berkontribusi terhadap upaya pencapaian pembangunan berkelanjutan melalui kebijakan yang lebih merata dan tepat sasaran.

2. Kajian Terdahulu

Beberapa penelitian terdahulu menunjukkan bahwa metode *K-Means Clustering* efektif dalam mengelompokkan wilayah berdasarkan berbagai indikator. Pada penelitian yang dilakukan oleh (Sandia & Dwidasmara, 2023) mengenai implementasi algoritma *K-Means Clustering* dalam mengklasifikasikan tingkat pembangunan perekonomian di Provinsi Bali menunjukkan bahwa metode ini berhasil mengklasifikasikan kabupaten/kota di Provinsi Bali berdasarkan indikator seperti PDRB per kapita, Indeks Pembangunan Manuasia (IPM), dan rata-rata lama sekolah. Penelitian ini juga berhasil mengidentifikasi ketimpangan yang terjadi antar wilayah serta memberikan rekomendasi berbasis data yang dapat djadikan pertimbangan dalam pengambilan kebijakan. Namun, fokus penelitian ini masih terbatas pada aspek ekonomi dan belum secara spesifik menganalisis indikator lain yang juga berpengaruh terhadap pembangunan wilayah.

Penelitian yang dilakukan oleh (Sukarno Wijaya et al., 2024) menunjukkan bahwa metode *K-Means Clustering* efektif digunakan untuk mengelompokkan wilayah berdasarkan tingkat kemiskinan dan pengangguran di Provinsi Jawa Timur. Hasil penelitian menunjukkan bahwa metode ini dapat mengkategorikan wilayah dengan tingkat kemiskinan rendah dan tinggi secara efektif. Namun, penelitian ini hanya menggunakan dua indikator utama, sehingga kurang mampu menggambarakan kondisi sosial secara menyeluruh.

Pada penelitian (Fajrianti et al., 2019) yang menggunakan metode *K-Means* dan analisis diskriminan untuk mengelompokkan desa-desa miskin di Kabupaten Pangkep juga menunjukkan hasil clusterisasi yang sangat akurat dengan tingkat akurasi tinggi mencapai 98,06%. Temuan ini menunjukkan bahwa metode *K-Mean Clustring* efektif dalam mengelompokkan wilayah berdasarkan indikator kemiskinan. Namun, keterbatasan ruang lingkup yang hanya mencakup satu kabupaten menjadi salah satu kelemahan dalam penelitian ini, karena membatasi generalisasi hasil penelitian ke wilayah lain yang memiliki karakteristik berbeda.

Penelitian lain oleh (Muttaqin & Zulkarnain, 2020) yang memanfaatkan metode *K-Means Clustering* untuk mengelompokkan kabupaten/kota di Indonesia berdasarkan indikator IPM berhasil mengidentifikasi tiga kategori wilayah pembangunan serta mengungkap adanya ketimpangan pembangunan yang signifikan antarwilayah. Hasil ini menunjukkan bahwa metode *K-Means Clustering* efektif dalam memetakan wilayah berdasrkan indikator IPM. Namun, penelitian ini hanya berfokus pada empat komponen IPM, sehingga kurang mampu merepresentasikan kondisi sosial ekonomi secara menyeluruh.

Berbeda dengan penelitian sebelumnya yang hanya menggunakan dua hingga tiga indikator saja, penelitian ini menggunakan kombinasi indikator sosial yang lebih komprehensif seperti, jumlah penduduk, pertumbuhan penduduk, kepadatan penduduk, tingkat kemiskinan, tingkat pengangguran terbuka (TPT), indeks pembangunan manusia (IPM), dan rata-rata lama sekolah (RLS). Penggunaan indikator sosial yang lebih luas memungkinkan analisis yang lebih mendalam terhadap berbagai aspek sosial yang mempengaruhi prioritas pembangunan. Selain itu, penelitian ini juga berupaya mengungkap karakteristik sosial utama yang membedakan antar klaster wilayah, sebagai dasar untuk merumuskan arah kebijakan pembangunan yang lebih merata dan sesuai dengan kondisi masing-masing kelompok daerah. Dengan pendekatan ini, wilayah-wilayah yang memerlukan perhatian dan tindak lanjut kebijakan dapat diidentifikasi secara lebih akurat. Metode ini tidak hanya memberikan gambaran yang lebih menyeluruh mengenai kondisi sosial suatu wilayah, tetapi juga mampu menangkap ketimpangan antarwilayah secara lebih mendalam. Sama seperti penelitian sebelumnya, penelitian ini menerapkan metode *K-Means Clustering* untuk memetakan wilayah berdasarkan beberapa indikator tertentu. Hasil dari penerapan metode ini

diharapkan dapat menjadi dasar yang kuat dalam perumusan kebijakan pembangunan yang lebih tepat sasaran, efektif, dan berkelanjutan.

Pembangunan

Menurut Todaro (2006) pembangunan pada hakikatnya merupakan suatu proses untuk melakukan perubahan pada aspek sosial dan ekonomi menuju arah yang lebih baik dan berkelanjutan (Wulandari & Tulis, 2022). Pembangunan memiliki tiga sasaran utama yang saling berkaitan, yaitu memperluas akses terhadap kebutuhan dasar (pangan, tempat tinggal dan kesehatan), meningkatkan taraf hidup masyarakat, serta memperluas pilihan ekonomi dan sosial bagi seluruh lapisan masyarakat (Afandi et al., 2022). Pembangunan dijadikan sebagai tolak ukur kemajuan dari suatu daerah, karena mencerminkan kesejahteraan masyarakat secara menyeluruh, tidak hanya dari aspek ekonomi, tetapi juga aspek sosial dan kehidupan masyarakat secara umum (Sriwati et al., 2024).

Clustering

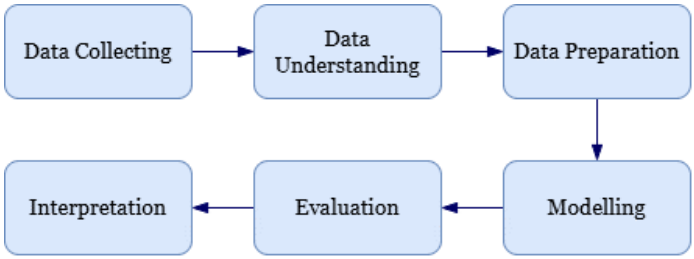
Clustering adalah metode pengelompokan data yang didasarkan pada model statitik, di mana data secara otomatis dikelompokkan ke dalam beberapa kelompok yang memiliki karakteristik serupa (Gormley et al., 2023). Teknik *Clustering* pada dasarnya adalah teknik analisis data yang digunkan untuk mengelompokkan data yang tidak memiliki label atau kategori tertentu (Oyewole & Thopil, 2023). Tujuan utamanya adalah untuk menemukan pola atau karakteristik tertentu dalam data, sehingga data dapat dikelompokkan ke dalam beberapa kelompok-kelompok yang memiliki kemiripan satu sama lain.

K-Means

K-Means merupakan salah satu metode clustering yang paling umum digunakan. Metode ini termasuk dalam teknik unsupervised learning, yaitu pendekatan *machine learning* yang digunakan untuk menganalisis data tanpa label atau kategori yang telah ditentukan sebelumnya (Sinaga & Yang, 2020). Metode ini bekerja dengan mengelompokkan data berdasarkan nilai rata-rata posisi objek pada setiap cluster. Pada penerapannya, metode ini memerlukan jumlah cluster yang perlu ditentukan terlebih dahulu untuk digunakan sebagai parameter awal (Az-Zahra & Wijayanto, 2024) . Pusat cluster awal akan dipilih secara acak dari data yang ada, kemudian selanjutnya akan diperbarui secara iteratif hingga mencapai hasil yang optimal (Ikotun et al., 2023). Algoritma ini sering digunakan karena memiliki waktu eksekusi yang cepat, implementasi yang sederhana, serta kemampuan untuk menangani data berukuran kecil secara efektif (Suraya et al., 2023).

3. Metodologi Penelitian

Secara garis besar, penelitian ini dilakukan melalui enam tahapan utama, yaitu meliputi *data collecting, data understanding, data preparation, modelling, evaluation, dan interpretation*. Gambaran dari tahapan penelitian ini disajikan dalam bentuk diagram yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Secara keseluruhan, tahapan-tahapan penelitian yang ditunjukkan pada Gambar 1 dapat dijelaskan sebagai berikut:

3.1 *Data Collecting*

Pada tahap ini, proses pengambilan data dilakukan dengan mengunjungi berbagai situs yang menyediakan data terkait indikator yang digunakan pada penelitian. Data yang digunakan pada penelitian ini merupakan data indikator sosial yang dianggap relevan dalam analisis wilayah, khususnya dalam mendukung pengambilan keputusan terkait derah yang menjadi prioritas pembangunan.

3.2 *Data Understanding*

Tahapan data *understanding* dilakukan dengan tujuan untuk memahami struktur dan karakteristik data yang akan digunakan dalam penelitian. Proses yang dilakukan mencakup deteksi *missing value* dan data *outlier*, identifikasi jenis atau tipe data, serta pengamatan tehadap distribusi dari masing-masing variabel. Tujuannya adalah untuk memberi gambaran yang lebih mendalam mengenai kondisi data secara keseluruhan sebelum digunakan untuk proses analisis lebih lanjut.

3.3 *Data Preparation*

Tahapan data preparation dilakukan dengan tujuan untuk mempersiapkan data agar siap digunakan dalam tahap pemodelan. Proses utama yang dilakukan dalam tahapan ini adalah normalisasi data, yaitu proses transformasi skala variabel agar berada di rentang yang seragam. Normalisasi dilakukan agar semua variabel memiliki bobot yang setara, sehingga tidak ada variabel tertentu yang mendominasi analisis. Dengan demikian, setiap variabel menjadi seimbang dalam proses clusterisasi dan hasil yang diperoleh lebih objektif dan representatif terhadap keseluruhan data. Pada penelitian ini, proses normaliasasi data dilakukan dengan menggunakan metode *MinmaxScaler*.

3.4 *Modelling*

Pada penelitian ini, metode *K-Means* digunakan untuk membangun model clusterisasi wilayah berdasarkan indikator sosial. Alasan pemilihan metode ini didasarkan pada kemampuannya yang mudah diimplementasikan, memiliki waktu eksekusi yang cepat, serta mampu menangani data dengan ukuran relatif kecil secara efektif. Adapun tahapan dari proses clusterisasi menggunakan metode *K-Means* dapat diuraikan sebagai berikut:

1. **Menentukan jumlah cluster (*k*)**
   Penentuan jumlah kluster merupakan langkah awal yang sangat penting dilakukan pada metode *K-Means Clustering*, karena pada penerapannya metode ini memerlukan jumlah cluster yang harus ditentukan terlebih dahulu untuk digunakan sebagai parameter awal proses clusterisasi. Pada penelitian ini penentuan jumlah cluster (*k*) dilakukan menggunakan *Elbow Method*.

2. **Menentukan pusat cluster awal (*centroid*)**
   Setelah jumlah cluster ditentukan, model akan secara acak memilih *K* data sebagai pusat awal cluster (*centroid*) yang nantinya digunakan sebagai acuan pengelompokan data.

3. **Menghitung jarak setiap data ke *centroid***
   Kemudian jarak masing-masing data ke *centroid* akan dihitung dengan rumus *Euclidean Distance*.

$$d(x_i, \mu_j) = \sqrt{(x_i - \mu_j)^2}$$

   Keterangan:
   - x_1 = Nilai *rate* atau voting dari data ke-*i* (misalnya indikator jumlah penduduk).
   - μ_j = Nilai *centroid* cluster ke-*j* saat ini.

4. **Mengelompokkan data berdasarkan jarak minimum ke *centroid***
   Setiap data akan dimasukkan ke dalam kelompok (*claster*) dengan berdasar pada jarak terdekat dari *centroid* yang telah dihitung pada langkah 3 sebelumnya.

5. **Memperbarui posisi *centroid***
   Setelah pengelompokan awal, centroid akan diperbarui dengan menghitung nilai rata-rata dari seluruh data yang telah berada dalam masing-masing kelompok (*cluster*) dengan rumus sebagai berikut:

$$C_k = \frac{1}{n_k} \sum_{i=1}^{n_k} d_i$$

   Keterangan:
   - C_k = centroid baru cluster ke-*k*

- n_k = jumlah data dalam cluster ke-k
- d_i = data ke-i dalam cluster tersebut

6. **Langkah 3-5 akan diulangi secara terus-menerus sampai tidak ada lagi perubahan yang signifikan dalam pembentukan cluster.**

3.5 *Evaluation*

Tahapan evaluasi dilakukan untuk mengukur kualitas cluster yang terbentuk setelah proses pemodelan. Pada penelitian ini, evaluasi dilakukan menggunakan *Silhouette Score*. Metode ini digunakan untuk menilai kualitas klaster dengan melihat seberapa baik data dikelompokkan ke dalam klaster yang sesuai dan seberapa jelas pemisahan antar klaster.

3.6 *Interpretation*

Tahapan interpretasi hasil dilakukan dengan memetakan hasil clusterisasi ke dalam peta wilayah Jawa Timur. Setiap kabupaten/kota yang telah dikelompokkan berdasarkan indikator sosial akan divisualisasikan menggunakan warna yang berbeda sesuai dengan cluster yang terbentuk. Pemetaan ini bertujuan untuk memberikan gambaran umum terkait distribusi wilayah dalam setiap cluster, sehingga dapat membantu mengidentifikasi wilayah prioritas pembangunan dengan lebih mudah.

4. Hasil dan Pembahasan

Pada bab ini, peneliti akan memaparkan hasil dari setiap tahapan penelitian, mulai dari persiapan dataset, pembuatan model, evaluasi dan implementasi hasil *clustering*.

4.1 Data Collecting

Data yang digunakan dalam penelitian ini merupakan data publik yang diambil dari situs resmi Badan Pusat Statistik (BPS) Provinsi Jawa Timur. Data yang diambil mencakup berbagai indikator sosial yang dianggap berpengaruh terhadap pembangunan suatu wilayah, yaitu jumlah penduduk, laju pertumbuhan penduduk, tingkat kemiskinan, Indek Pembangunan Manusia (IPM), Tingkat Pengangguran Terbuka (TPT), dan rata-rata lama sekolah pada tingkat kabupaten/kota di provinsi Jawa Timur pada tahun 2024. Gambaran umum dari dataset yang digunakan dapat dilihat pada Gambar 2.

	Kabupaten/Kota	Jumlah Penduduk	Laju Pertumbuhan Penduduk	Kepadatan Penduduk (Km2)	Tingkat Kemiskinan	IPM	TPT	RLS
0	Pacitan	588.6	0.11	411	13.08	71.49	1.56	11.920
1	Ponorogo	962.9	0.38	679	9.11	73.70	4.19	11.850
2	Trenggalek	744.5	0.49	596	10.50	72.47	3.90	12.020
3	Tulungagung	1113.9	0.58	973	6.28	75.13	4.12	13.175
4	Blitar	1263.7	0.86	724	8.16	73.44	4.77	11.855

Gambar 2. Data Indikator Clustering

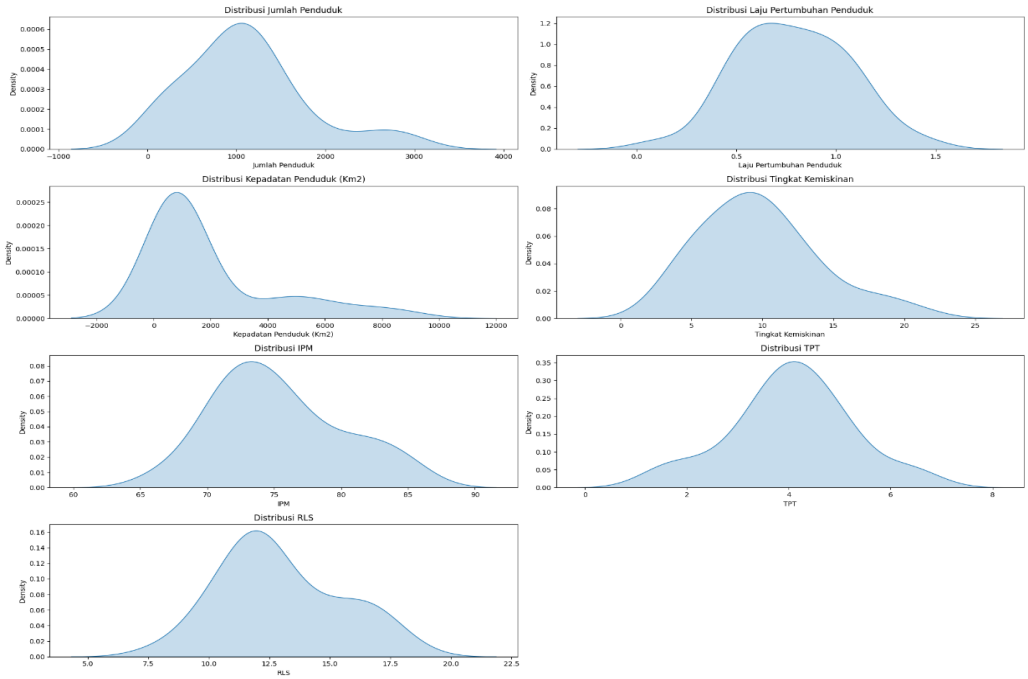
4.2 Data Understanding

Pada tahap ini, diketahui bahwa data yang digunakan tidak memiliki nilai kosong (missing value), sebagaimana ditunjukkan pada Gambar 3.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38 entries, 0 to 37
Data columns (total 8 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Kabupaten/Kota              38 non-null     object
 1   Jumlah Penduduk             38 non-null     float64
 2   Laju Pertumbuhan Penduduk   38 non-null     float64
 3   Kepadatan Penduduk (Km2)    38 non-null     int64
 4   Tingkat Kemiskinan          38 non-null     float64
 5   IPM                         38 non-null     float64
 6   TPT                         38 non-null     float64
 7   RLS                         38 non-null     float64
dtypes: float64(6), int64(1), object(1)
memory usage: 2.5+ KB
```

Gambar 3. Identifikasi *Missing Value*

Namun, pada tahap analisis distribusi data diketahui bahwa terdapat beberapa fitur yang memiliki distribusi tidak seimbang. Hasil visualisasi distribusi yang ditampilkan pada Gambar 4 menunjukkan bahwa sebagian besar variabel, seperti laju pertumbuhan penduduk, tingkat kemiskinan, IPM, dan tingkat pengangguran terbuka memiliki pola distribusi mendekati normal, meskipun terdapat beberapa data yang menunjukkan skew. Disisi lain, variabel seperti jumlah penduduk, kepadatan penduduk, dan rata-rata lama sekolah (RLS) menunjukkan distribusi yang melenceng ke kanan (*right-skewed*), yang mengindikasikan adanya ketimpangan nilai pada beberapa wilayah. Kondisi ini menunjukkan bahwa pola distribusi data perlu diproses lebih lanjut sebelum digunakan dalam proses klasterisasi, guna meminimalkan dominasi nilai-nilai ekstrem yang dapat memengaruhi hasil pengelompokan.



Gambar 4. Plot Distribusi Data

4.3 Data Preparation

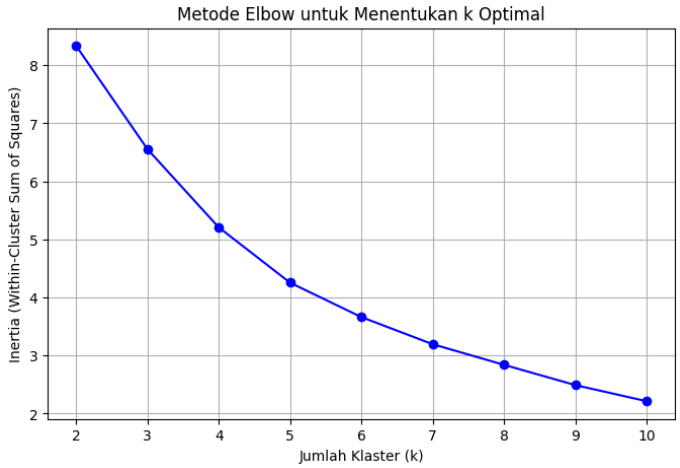
Berdasarkan hasil analisis ditribusi data yang ditampilan pada Gambar 4, diketahui bahwa data yang akan digunakan dalam proses klasterisasi memiliki distribusi yang tidak seimbang. Untuk mengatasi hal tersebut, pada tahap ini dilakukan proses normalisasi data menggunakan metode *MinMaxScaler*, dimana data diubah dalam skala rentang 0 hingga 1. Tujuannya untuk menghindari dominasi fitur yang memiliki nilai tinggi dan memastikan bahwa setiap variabel memiliki kontribusi yang seimbang dalam proses klasterisasi data. Normalisasi penting untuk dilakukann agar hasil klasterisasi lebih akurat dan representatif terhadap karakteristik setiap wilayah. Hasil dari proses normalisasi data akan terlihat seperti pada Gambar 5.

	Kabupaten/Kota	Jumlah Penduduk	Laju Pertumbuhan Penduduk	Kepadatan Penduduk (Km2)	Tingkat Kemiskinan	IPM	TPT	RLS
0	Pacitan	0.162034	0.000000	0.000000	0.563872	0.265442	0.000000	0.387518
1	Ponorogo	0.296452	0.204545	0.032340	0.340461	0.388425	0.533469	0.380745
2	Trenggalek	0.218021	0.287879	0.022324	0.418683	0.319978	0.474645	0.397194
3	Tulungagung	0.350679	0.356061	0.067817	0.181204	0.468002	0.519270	0.508950
4	Blitar	0.404475	0.568182	0.037770	0.287001	0.373957	0.651116	0.381229

Gambar 5. Hasil Normalisasi Data

4.4 Modelling

Sebelum masuk kedalam tahap klasterisasi, dilakukan penentuan nilai K optimal yang akan digunakan sebagai parameter awal dalam proses klasifikasi menggunakan *Elbow Method*. Dari hasil visualisasi yang ditunjukkan pada Gambar 6, diketahui bahwa nilai *K=2* merupakan jumlah klaster yang paling optimal. Hal ini diperjelas oleh posisi titik siku pada grafik yang menunjukkan bahwa penurunan nilai inertia (*within-cluster sum of squares*) mulai melambat secara signifikan setelah *K = 2*. Oleh karena itu, nilai *K=2* dipilih sebagai parameter yang digunakan dalam proses klasterisasi.



Gambar 6. Grafik *Elbow Method*

Setelah ditentukan bahwa *K=2* merupakan jumlah klaster optimal, proses klasterisasi dilanjutkan dengan menerapkan algoritma *K-Means clustering* untuk mengelompokkan wilayah berdasarkan indikator-indikator sosial yang telah ditentukan sebelumnya. Data yang digunakan dalam proses klasterisasi merupakan dataset yang telah melalui proses normalisasi sebelumnya. Hasil klasterisasi menunjukkan bahwa wilayah terbagi menjadi dua klaster dengan distribusi yang berbeda. Klaster 0 terdiri dari 11 wilayah, sedangkan Klaster 1 mencakup 27 wilayah. Perbedaan distribusi setiap klaster mencerminkan adanya perbedaan karakteristik sosial antar wilayah di Provinsi Jawa Timur. Daftar wilayah dari masing-masing klaster dapat dilihat pada Tabel 1.

Tabel 1. Daftar Wilayah Berdasarkan *Cluster*

C0	C1
Kab. Sidoarjo	Kab. Pacitan
Kab. Gresik	Kab. Ponorogo
Kota Kediri	Kab. Trenggalek
Kota Blitar	Kab. Tulungagung
Kota Malang	Kab. Blitar
Kota Probolinggo	Kab. Kediri
Kota Pasuruan	Kab. Malang
Kota Mojokerto	Kab. Lumajang
Kota Madiun	Kab. Jember
Kota Surabaya	Kab. Banyuwangi
Kota Batu	Kab. Bondowoso
	Kab. Situbondo
	Kab. Probolinggo
	Kab. Pasuruan
	Kab. Mojokerto
	Kab. Jombang
	Kab. Nganjuk
	Kab. Madiun
	Kab. Magetan
	Kab. Ngawi
	Kab. Bojonegoro
	Kab. Tuban
	Kab. Lamongan
	Kab. Bangkalan
	Kab. Sampang
	Kab. Pamekasan
	Kab. Sumenep

Berdasarkan hasil klasterisasi, wilayah-wilayah yang termasuk kedalam Klaster 0 memiliki karakteristik sosial yang lebih baik dibandingkan dengan wilayah-wilayah pada Klaster 1. Sebagaimana yang ditampilkan pada Gambar 7, wilayah yang termasuk dalam Klaster 0 memiliki rata-rata jumlah penduduk sebesar 801 ribu jiwa, dengan kepadatan yang cukup tinggi yaitu 4859 jiwa/km^2, serta laju pertumbuhan penduduk yang tinggi yaitu sebesar 1,02%. Namun, wilayah dalam klaster ini menunjukkan kondisi sosial yang lebih baik, ditandai dengan tingkat kemiskinan yang rendah yaitu 5,59%, Indeks Pembangunan Manusia (IPM) yang tinggi sebesar 81,54%, dan rata-rata lama sekolah (RLS) sebesar 16,24 tahun.

Sebaliknya, wilayah-wilayah yang termasuk dalam Klaster 1 memiliki rata-rata jumlah pendudukan yang lebih besar, yaitu sekitar 1,22 juta jiwa. Namun, kepadatan dan laju pertumbuhan penduduk di wilayah Klaster 0 jauh lebih rendah dibandingkan dengan wilayah pada Klaster 1, yaitu masing-masing sebesar 770 jiwa/km^2 dan 0,69%. Dari segi indikator sosial lainnya, wilayah-wilayah pada Kaster 1 menunjukkan tingkat kemiskinan yang lebih tinggi sebesar 11,49%, Indeks Pembangunan Manusia (IPM) yang lebih rendah sebesar 72,77%, dan rata-rata lama sekolah (RLS) yang lebih rendah juga, yaitu 11,61 tahun.

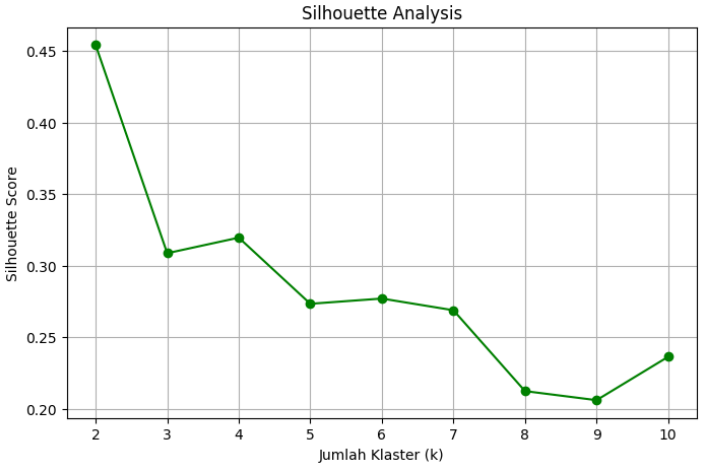
Hasil ini menunjukkan bahwa, wilayah yang termasuk dalam Klaster 0 cenderung memiliki keadaan sosial yang lebih maju, terutama ditinjau dari tiga indikator utama, yaitu tingkat kemiskinan, Indeks Pembangunan Manusia (IPM), dan rata-rata lama sekolah (RLS). Ketimpangan nilai pada tiga indikator ini menjadi penanda utama yang menentukan bahwa wilayah yang termasuk kedalam Klaster 1 merupakan daerah yang menjadi prioritas pembangunan, sementara wilayah yang termasuk dalam Klaster 0 berada dalam kondisi sosial yang relatif baik dan tidak menjadi daerah prioritas utama pembangunan. Dengan demikian, hasil dari klasterisasi ini dapat digunakan sebagai dasar dalam perumusan kebijakan pembangunan daerah. Pemerintah dapat memfokuskan anggaran dan program pembangunan pada wilayah yang termasuk dalam Klaster 1 untuk memperbaiki indikator sosial yang masih tertinggal. Selain itu, pemetaan ini juga memungkinkan adanya pendekatan pembangunan yang lebih adaptif dan berbasis data, mengingat setiap kluster menunjukkan karakteristik sosial yang berbeda secara signifikan.

Jumlah Penduduk	Laju Pertumbuhan Penduduk	Kepadatan Penduduk (Km2)	Tingkat Kemiskinan	IPM	TPT	RLS	Cluster
801.281818	1.016364	4859.818182	5.59000	81.540000	4.884545	16.237273	0.0
1222.244444	0.687037	770.555556	11.49037	72.772593	3.697778	11.606667	1.0

Gambar 7. Rata-rata Nilai Indikator dari Setiap *Cluster*

4.5 Evaluasi

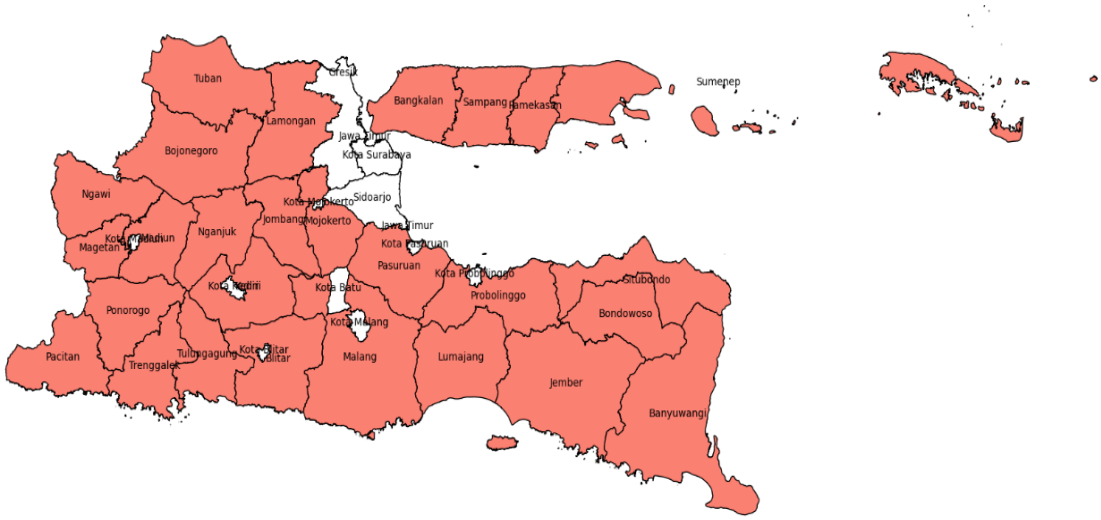
Untuk memastikan bahwa hasil klasterisasi yang dilakukan sebelumnya berjalan dengan optimal, dilakukan evaluasi menggunakan metode *Silhouette*. Metode ini digunakan untuk mengukur seberapa baik objek-objek dikelompokkan ke dalam klaster yang tepat, dimana semakin tinggi skor yang dihasilkan, maka semakin baik pemisahan antar klaster yang terbentuk. Berdasarkan hasil visualisasi yang ditunjukkan pada Gambar 8, nilai *Silhouette Score* tertinggi diperoleh pada saat jumlah *K=2*, yaitu sebesar 0,45. Temuan ini sejalan dengan hasil dari metode *Elbow* yang telah dilakukan sebelumnya. Konsistensi ini menguatkan asumsi bahwa pemilihan *K=2* merupakan keputusan paling tepat, karena menghasilkan pemisahan klaster yang konsisten berdasarkan indikator sosial yang digunakan.



Gambar 8. Grafik *Silhouette Score*

4.6 Interpretasi Hasil

Data hasil klasterisasi kemudian divisualisasikan menggunakan *GeoPandas* untuk melihat sebaran masing-masing klaster. Sebagaimana yang terlihat pada Gambar 9, wilayah yang termasuk dalam prioritas pembangunan (Klaster 1) ditandai dengan warna merah, sedangkan untuk wilayah yang bukan prioritas pembangunan (Klaster 0) ditandai dengan warna putih. Berdasarkan hasil visualisasi, terlihat bahwa sebagian besar wilayah yang menjadi prioritas merupakan daerah kabupaten, sementara wilayah yang termasuk non-prioritas umumnya merupakan daerah perkotaan. Hasil ini menunjukkan adanya ketimpangan pembangunan antara wilayah urban dan rural, dimana wilayah rural (pedesaan/kabupaten) cenderung menjadi daerah prioritas pembangunan karena rendahnya nilai capaian pada beberapa indikator sosial seperti tingkat pendidikan (RLS), IPM, dan tingkat kemiskinan.



Gambar 9. Peta Daerah Prioritas Pembangunan

Meskipun hasil penelitian menunjukkan bahwa metode K-Means Clustering mampu mengelompokkan wilayah secara objektif berdasarkan indikator sosial, terdapat beberapa keterbatasan yang perlu diperhatikan. Pertama, data yang digunakan dalam penelitian ini masih berfokus pada indikator sosial dan belum mencakup aspek lain seperti akses terhadap infrastruktur, layanan kesehatan, maupun kualitas lingkungan yang juga berpengaruh dalam menentukan prioritas pembangunan. Kedua, pendekatan yang digunakan bersifat statis karena hanya mengandalkan data dalam satu periode waktu, sehingga belum mampu menangkap dinamika sosial yang dapat berubah dari tahun ke tahun. Ketiga, metode *K-Means Clustering* sangat bergantung pada penentuan jumlah klaster (nilai *K*) di awal, dan memiliki sensitivitas yang tinggi terhadap keberadaaan data outlier, meskipun proses normalisasi telah dilakukan. Hal ini dapat menyebabkan bias dalam interpretasi, terutama jika data yang digunakan memiliki ditribusi yang tidak seimbang atau terdapat nilai-nilai yang berbeda secara signifikan.

Oleh karena itu, untuk penelitian selanjutnya disarankan untuk memperluas cakupan indikator yang digunakan, dengan memasukkan variabel-variabel yang lebih beragam dari berbagai aspek, seperti akses terhadap infrastruktur, layanan kesehatan, dan kualitas lingkungan, guna menghasilkan klasifikasi wilayah yang lebih komprehesif. Selain itu, penggunaan data time series juga dapat dipertimbangkan agar analisis tidak hanya menggambarkan kondisi pada satu periode tertentu, tetapi juga mampu menangkap dinamika perubahan sosial antarwilayah dari waktu ke waktu. Terakhir, dapat dilakukan penerapan metode klasterisasi lain seperti DBSCAN atau *Hierarchical Clustering* guna memperoleh model pengelompokan wilayah yang paling representatif.

5. Kesimpulan

Penelitian ini berhasil menerapkan pendekatan data mining untuk mengidentifikasi wilayah-wilayah prioritas pembangunan berdasarkan sejumlah indikator sosial, yaitu jumlah penduduk, laju pertumbuhan penduduk, kepadatan penduduk, IPM, tingkat pengangguran, dan rata-rata lama sekolah (RLS) dengan cukup baik. Dengan menggunakan metode *K-Means Clustering*, proses klasterisasi menghasilkan dua kategori wilayah, yakni wilayah prioritas pembangunan (Klaster 1) dan wilayah non-prioritas (Klaster 0). Dari hasil klasterisasi, diketahui bahwa sebagian besar wilayah yang menjadi prioritas pembangunan merupakan daerah kabupaten, sementara wilayah yang termasuk non-prioritas umumnya merupakan daerah perkotaan. Hasil evaluasi menggunakan *Silhoutte Score* sebesar 0,45 menunjukkan bahwa model menghasilkan pemisahan klaster yang cukup baik dan konsisten berdasarkan indikator sosial yang digunakan. Nilai ini mengindikasikan bahwa anggota dari masing-masing klaster memiliki kemiripan yang relatif tinggi satu sama lain, dan berbeda secara signifikan dengan anggota klaster lainnya.

Meskipun hasil penelitian ini memberikan gambaran awal yang cukup informatif terkait pengelompokan wilayah prioritas pembangunan, terdapat beberapa keterbatasan yang perlu diperhatikan. Data yang digunakan dalam penelitian ini relatif kecil dan terbatas pada satu provinsi, sehingga ada kemungkinan terjadinya bias dalam proses klasterisasi. Selain itu, indikator yang digunakan masih terbatas pada aspek sosial, tanpa mempertimbangkan faktor lain seperti kondisi ekonomi dan lingkungan yang juga berperan dalam menentukan prioritas pembangunan. Sehingga, hasil pemetaan belum sepenuhnya merepresentasikan kondisi wilayah secara keseluruhan. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas cakupan indikator dengan memasukkan variabel tambahan seperti infrastruktur, layanan kesehatan, dan kualitas lingkungan, serta mempertimbangkan penggunaan data *time series* untuk menangkap dinamika perubahan sosial antarwilayah secara lebih menyeluruh, sehingga hasil klasifikasi yang diperoleh dapat mencerminkan kondisi dan kebutuhan pembangunan secara lebih akurat dan berkelanjutan.

Meskipun demikian, penelitian ini memberikan kontribusi penting dalam pengembangan data mining dan analisis spasial untuk perencanaan pembangunan wilayah. Penerapan metode *K-Means Clustering* terbukti dapat menjadi alat bantu yang efektif dalam proses pengambilan keputusan, khususnya dalam menentukan wilayah prioritas pembangunan berbasis data. Dengan demikian, pendekatan ini tidak hanya memperkuat dasar pengambilan kebijakan secara objektif, tetapi juga membuka peluang untuk pengembangan lebih lanjut dengan memperluas cakupan wilayah dan menambah indikator yang lebih beragam, sehingga mendukung perencanaan pembangunan yang lebih komprehensif, adaptif, dan tepat sasaran.

6. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, bimbingan, dan bantuan dalam proses pelaksanaan penelitian ini.

7. Pernyataan Penulis

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menyatakan bahwa data dan makalah bebas dari plagiarisme serta penulis bertanggung jawab secara penuh atas keaslian artikel.

Bibliografi

Amalia, A. F., Sadik, J., Azzahra, R. D., & Mubarok, I. T. (2024). *Analisis Pengaruh Ketimpangan Pendapatan, Pertumbuhan Ekonomi, dan Jumlah Penduduk Terhadap Kemiskinan di Jawa Timur Tahun 2023*. *5*(1), 140–151.

Anwar, A. A., Rorong, I. P. F., & Tolosang, K. D. (2023). Analisis Ketimpangan Pembangunan Antar Wilayah Di Kabupaten/Kota Provinsi Sulawesi Utara. *Jurnal Berkala Ilmiah Efisiensi*, *23*(6), 85–96.

Az-Zahra, A., & Wijayanto, A. W. (2024). Tinjauan Kesejahteraan di Daerah Perbatasan Republik Indonesia Tahun 2021: Penerapan Analisis Klaster K-Means dan Hierarki. *Jurnal Sistem Dan Teknologi Informasi*, *12*(1), 55–64. https://doi.org/10.26418/justin.v12i1.69040

Fajrianti, F., Bustan, M. N., & Tiro, M. A. (2019). Penggunaan Analisis Cluster K-Means dan Analisis Diskriminan Dalam Pengelompokan Desa Miskin di Kabupaten Pangkep. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *1*(2), 7. https://doi.org/10.35580/variansiunm9355

Gormley, I. C., Murphy, T. B., & Raftery, A. E. (2023). Model-Based Clustering. In *Annual Review of Statistics and Its Application* (Vol. 10, pp. 573–595). Annual Reviews Inc. https://doi.org/10.1146/annurev-statistics-033121-115326

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Muhammad Fakhrur Rodzi. (2023). Pembangunan Infrastruktur Dan Pemerataan Ekonomi Di Indonesia. *Jurnal Masyarakat Dan Desa*, *3*(2), 151–163. https://doi.org/10.47431/jmd.v3i2.353

Muttaqin, M. F. J., & Zulkarnain. (2020). Cluster Analysis Using K-Means Method to Classify Indonesia Regency/City based on Human Development Index Indicator. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3400934.3400951

Nurvia Handayani, N. H. (2024). *Pengaruh Ketimpangan Pendapatan, Tingkat Pendidikan dan Kemiskinan Terhadap Pertumbuhan Ekonomi Di Indonesia Nurvia*. *4*, 112–124.

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7). https://doi.org/10.1007/s10462-022-10325-y

Rinaldi, M. (2020). *Pendidikan sebagai Pilar Kesejahteraan : Menghubungkan Pendidikan dengan Kemajuan Sosial dan Ekonomi*. *08*(1).

Sandia, I. W. W. K., & Dwidasmara, I. B. G. (2023). Implementasi Algoritma K-Means Clustering dalam Penentuan Klasifikasi Tingkat Pembangunan Perekonomian di Provinsi Bali. *Jurnal Nasional Teknologi Dan Aplikasi*, *1*(2), 761–769.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Sriwati, E., Setiawati, B., & Tahir, N. (2024). Peran pemerintah daerah dalam pembangunan infrastruktur. *Jurnal KIMAP: Kajian Ilmiah Mahasiswa Administrasi Publik*, *5*(1), 104–116. https://journal.unismuh.ac.id/index.php/kimap/article/view/14058

Sukarno Wijaya, N., Jajuli, M., & Arif Dermawan, B. (2024). Penerapan Algoritma K-Means Clustering Dalam Menentukan Daerah Prioritas Penanganan Kemiskinan Di Wilayah Jawa Timur. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(4), 7579–7584. https://doi.org/10.36040/jati.v8i4.10248

Suraya, S., Sholeh, M., & Lestari, U. (2023). Evaluation of Data Clustering Accuracy using K-Means Algorithm. *International Journal of Multidisciplinary Approach Research and Science*, *2*(01). https://doi.org/10.59653/ijmars.v2i01.504

Syed Agung Afandi, Muslim Afandi, R. E. (2022). Pengantar Teori Pembangunan. In N. H. Afandi (Ed.), *CV. Bintang Semesta Media* (1st ed.). Percetakan Bintang. http://repository.ut.ac.id/4601/

Wulandari, S., & Tulis, R. S. (2022). Prisip Manajemen Dalam Proses Pembagunan Infastruktur Di Kabupaten Katingan (Studi Di Desa Tumbang Lahang) Jurnal Administrasi Publik (JAP). *Jurnal Administrasi Publik (JAP)*, *8*(2), 148–161.