JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 134 – 146
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Klasterisasi Jumlah Penduduk Provinsi Jawa Timur Tahun 2021-2023 Menggunakan Algoritma K-Means

Risqi Pradana Aryanto [(1)*], Agung Nilogiri [(2)], Ari Eko Wardoyo [(3)]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember, Jember
e-mail : riskipradana221001@gmail.com, {agungnilogiri,arieko}@unmuhjember.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 9 Februari 2024, direvisi 16 Maret 2024, diterima 17 Maret 2024, dan dipublikasikan 25 Mei 2024.

Abstract

Understanding the population data of a region is crucial for policy development and strategic planning. East Java Province, the second-largest province in Indonesia, has undergone significant population growth from 2021 to 2023. Uneven growth poses challenges in resource and infrastructure management. The K-Means algorithm clusters population data into several groups based on specific characteristics. The Elbow *method is used to determine the optimal number of clusters, ensuring the accuracy of the analysis. This research aims to analyze and cluster the population distribution in each city in East Java Province, providing a more detailed and accurate depiction. The research findings reveal three significant clusters. Cluster 0 includes 21 towns, Cluster 1 comprises 4, and Cluster 2 encompasses 13. These findings have important implications for targeted development policy formulation at the city level in East Java Province. Additionally, this study contributes to the development of demographic analysis and population management, using valid methods and consistent results between RapidMiner and manual calculations. In conclusion, this research provides a solid foundation for more effective development policy formulation in East Java Province, offering essential information for sustainable population management.*

Keywords: Population Distribution, Clustering, East Java, K-Means, Elbow Method, Data Mining

Abstrak

Pemahaman terhadap data populasi suatu wilayah menjadi hal yang sangat penting untuk pengembangan kebijakan dan perencanaan strategis. Provinsi Jawa Timur, sebagai provinsi kedua terbesar di Indonesia, mengalami perkembangan penduduk signifikan dari tahun 2021 hingga 2023. Pertumbuhan yang tidak merata menimbulkan tantangan dalam pengelolaan sumber daya dan infrastruktur. Algoritma K-Means digunakan sebagai solusi untuk mengelompokkan data jumlah penduduk ke dalam beberapa kelompok berdasarkan karakteristik tertentu. Metode Elbow digunakan untuk menentukan jumlah klaster optimal, memastikan keakuratan analisis. Tujuan penelitian ini adalah menganalisis dan mengklasterisasi sebaran penduduk di setiap kota di Provinsi Jawa Timur, memberikan gambaran yang lebih rinci dan akurat. Hasil penelitian menunjukkan adanya tiga klaster yang signifikan. *Cluster* 0 mencakup 21 kota, *Cluster* 1 mencakup 4 kota, dan *Cluster* 2 mencakup 13 kota. Temuan ini memiliki implikasi penting untuk perumusan kebijakan pembangunan yang lebih tepat sasaran di tingkat kota di Provinsi Jawa Timur. Selain itu, penelitian ini memberikan kontribusi pada pengembangan bidang analisis demografis dan pengelolaan populasi, dengan metode yang valid dan hasil yang konsisten antara penggunaan RapidMiner dan perhitungan manual. Kesimpulannya, penelitian ini menyediakan landasan yang kuat bagi perumusan kebijakan pembangunan yang lebih efektif di Provinsi Jawa Timur, memberikan informasi yang diperlukan untuk pengelolaan populasi yang berkelanjutan.

Kata Kunci: Sebaran Penduduk, Pengelompokan, Jawa Timur, K-Means, Metode Elbow, *Data Mining*

1. PENDAHULUAN

Pada era modern ini, pemahaman yang mendalam terhadap data populasi suatu wilayah menjadi sangat krusial untuk pengembangan kebijakan dan perencanaan strategis. Provinsi Jawa Timur, yang terletak di bagian timur Pulau Jawa, Indonesia, adalah provinsi dengan jumlah penduduk terbesar kedua di negara ini. Provinsi Jawa Timur merupakan salah satu wilayah yang terus mengalami perkembangan penduduk yang signifikan dari tahun 2021 hingga 2023. Fenomena ini menjadi sorotan karena adanya kebutuhan untuk memahami dan mengelola distribusi penduduk secara efektif. Dengan lebih dari 40 juta jiwa, Jawa Timur menjadi rumah bagi berbagai kelompok etnis, budaya, dan ekonomi.

Pertumbuhan penduduk di Jawa Timur memiliki pola yang unik dan beragam. Beberapa wilayah mengalami pertumbuhan penduduk yang pesat, sementara wilayah lainnya mengalami pertumbuhan yang lebih lambat atau bahkan penurunan jumlah penduduk. Faktor-faktor seperti urbanisasi, migrasi, dan perubahan sosial-ekonomi berkontribusi terhadap dinamika ini. Pertumbuhan penduduk yang tidak merata di berbagai kota di Provinsi Jawa Timur menimbulkan tantangan tersendiri dalam mengelola sumber daya dan pembangunan infrastruktur. Oleh karena itu, diperlukan suatu pendekatan analisis yang dapat mengidentifikasi pola dan karakteristik dalam distribusi jumlah penduduk.

Meskipun data jumlah penduduk tersedia, masih terdapat kendala dalam memahami pola sebaran penduduk pada tingkat kota di Provinsi Jawa Timur. Ketidakmampuan dalam mengidentifikasi kelompok atau pola tertentu dapat menghambat upaya perencanaan pembangunan yang efisien. Oleh karena itu, perlu dilakukan analisis lebih lanjut untuk mengatasi masalah ini. Maka digunakan algoritma K-Means sebagai metode analisis untuk mengklasterisasi jumlah penduduk pada setiap kota di Provinsi Jawa Timur. K-Means merupakan algoritma klasterisasi yang dapat mengelompokkan data ke dalam beberapa kelompok berdasarkan kesamaan karakteristik tertentu (Nofiar et al., 2019). Penggunaan algoritma ini diharapkan dapat memberikan gambaran yang jelas mengenai pola sebaran penduduk di wilayah tersebut.

K-Means bekerja dengan mengelompokkan data ke dalam k kelompok (klaster) yang ditentukan sebelumnya, dengan meminimalkan variasi atau jarak antara titik data dalam satu klaster. Pendekatan ini sangat relevan dalam konteks analisis sebaran penduduk, karena dapat membantu mengidentifikasi pola sebaran yang mungkin tidak terlihat secara langsung (Talakua et al., 2017). Dengan menggunakan K-Means, dapat mengelompokkan kota-kota berdasarkan tingkat persebaran jumlah penduduk dari rentang tahun 2021 hingga tahun 2023. Selain itu, untuk menentukan jumlah klaster yang optimal, penelitian ini akan menerapkan metode siku (Elbow *method*). Metode ini melibatkan pengujian berbagai jumlah klaster dan mengamati tingkat variasi yang dijelaskan oleh model terhadap jumlah klaster tersebut.

Metode Elbow atau Elbow *method* merupakan suatu teknik yang dimanfaatkan untuk mengidentifikasi jumlah klaster yang optimal dalam proses klastering, terutama diterapkan pada algoritma K-Means. Konsep inti dari metode ini adalah dengan mengamati penurunan varians di dalam setiap klaster seiring dengan variasi jumlah klaster yang berbeda (Syahfitri et al., 2023). Metode Elbow bekerja dengan mengukur varians atau dispersi data dalam setiap klaster saat jumlah klaster berubah-ubah. Proses ini melibatkan iterasi dengan berbagai jumlah klaster, dan pada setiap iterasi, varians dihitung untuk setiap klaster. Hasilnya kemudian dianalisis untuk mengidentifikasi titik di mana penurunan varians menjadi kurang signifikan, menandakan bahwa penambahan klaster tidak lagi memberikan keuntungan substansial dalam mengurangi dispersi data (Fahrozi et al., 2023). Dengan menggunakan metode Elbow, peneliti mengambil keputusan yang lebih informasional dan berbasis data dalam menentukan jumlah klaster optimal untuk suatu *dataset* tertentu. Teknik ini menjadi kunci dalam memahami struktur data dan mengoptimalkan kinerja algoritma klastering, khususnya algoritma K-Means, untuk mencapai segmentasi yang optimal dan lebih bermakna.

Penelitian terkait dengan penggunaan K-Means antara lain “Penggunaan K-Means untuk Klasterisasi Penetapan Instruktur Diklat pada PT PLN (Persero) UDIKLAT Jakarta” (Budiana et al., 2019), “Penggunaan K-Means *Clustering* untuk Klasterisasi Tingkat Kehadiran Dosen” (Virgo et al., 2020), dan “Penggunaan K-Means untuk Klastering Sayuran Unggulan” (Harahap et al., 2022). Namun pada penelitian Harahap tidak digunakan metode lain untuk menentukan jumlah klaster atau k sehingga klaster yang dihasilkan kurang begitu optimal. Penelitian lainnya yang berkaitan dengan penggunaan K-Means digunakan untuk klasterisasi siswa yang berprestasi (Dewi et al., 2022). Pada penelitian ini, jumlah kluster ditentukan menggunakan metode Davies Bouldin sehingga dapat menghasilkan jumlah kluster yang optimal. Penelitian lainnya terkait penggunaan metode Elbow yaitu metode ini digunakan untuk *clustering* pemerataan bantuan sosial di Kabupaten Bojonegoro (Fitriyah et al., 2023) dan algoritma K-Means *Clustering* Metode Elbow digunakan untuk menganalisa motivasi pengunjung Festival Halal JHF (Wicaksana et al., 2023).

Tujuan utama dari penelitian ini adalah untuk menganalisis dan mengklasterisasi sebaran jumlah penduduk pada setiap kota di Provinsi Jawa Timur menggunakan algoritma K-Means. Metode K-Means akan diaplikasikan dengan menggunakan metode Elbow untuk menentukan jumlah klaster yang optimal. Maka dari itu, penelitian ini bertujuan untuk mendapatkan informasi yang lebih rinci dan akurat mengenai pola sebaran penduduk di tingkat kota. Diharapkan penelitian ini bisa berkontribusi terhadap pengembangan bidang analisis demografis dan pengelolaan populasi serta menjadi landasan yang kuat bagi perumusan kebijakan pembangunan yang lebih efektif di Provinsi Jawa Timur.

2. METODE PENELITIAN

Metodologi penelitian merupakan landasan yang digunakan oleh peneliti untuk melaksanakan penelitian. Landasan ini mencakup serangkaian langkah dalam pengelolaan data, dimulai dari analisis kebutuhan hingga pemahaman terhadap hasil penelitian. Proses tersebut dijelaskan melalui beberapa tahapan yang terstruktur dan sistematis, sebagaimana yang ditampilkan dalam Gambar 1.

Gambar 1 Alur Kerja Penelitian

Berdasarkan Gambar 1 yang menggambarkan alur kerja penelitian, penelitian ini dimulai dengan pengumpulan data dan mengolah data yang akan menjadi data utama. Selanjutnya, dilakukan proses analisa data yang melibatkan tahap pembersihan data, dan transformasi data untuk menggabungkan serta mengubah format data agar sesuai dengan kebutuhan. Tahap selanjutnya, yaitu *data mining*, pada tahap ini, peneliti menerapkan metode K-Means *clustering* untuk implementasi *data mining*. Langkah terakhir melibatkan pengujian hasil, yang dilakukan dengan perhitungan manual dan menggunakan aplikasi RapidMiner.

2.1 Pengumpulan Data

Pengumpulan data merupakan suatu proses untuk mendapatkan informasi atau data yang diperlukan untuk menjawab pertanyaan dari penelitian yang diajukan. Proses ini melibatkan kegiatan sistematis dalam mencari dan menghimpun data secara langsung dari sumbernya baik data di lapangan maupun data dari sumber literatur lainnya (Anufia & Alhamid, 2019). Tujuan utama dari pengumpulan data adalah untuk memperoleh informasi yang relevan dan fakta yang dapat memberikan pemahaman yang mendalam terhadap permasalahan penelitian yang tengah diteliti.

2.2 Mengolah Data

Pengolahan data merupakan serangkaian kegiatan, yaitu pengumpulan, pemrosesan, dan analisis data. Proses pengolahan data bertujuan untuk menghasilkan informasi yang akan digunakan dalam penelitian. Hasil dari proses ini dapat berupa laporan, grafik, atau tabel yang memberikan representasi visual atau naratif dari temuan dan hasil analisis (Nawassyarif et al., 2020).

2.3 Menganalisa Data

Menganalisa data adalah langkah yang dilakukan untuk mengubah data yang telah dikumpulkan dan dibersihkan menjadi informasi yang memiliki nilai dan bermanfaat. Tujuan utama dari analisis data adalah untuk mengidentifikasi tren, pola, serta hubungan yang dapat ditemukan antara berbagai data set yang berbeda. Melalui proses ini, peneliti mampu mengidentifikasi dan menemukan karakteristik dari data yang telah terkumpul (Herviany et al., 2021).

2.4 *Data Mining*

Data mining merupakan suatu proses yang dilakukan untuk menemukan informasi atau pola menarik dalam *dataset* yang telah dipilih. Dalam melaksanakan proses ini, digunakan teknik atau metode khusus yang dapat mencakup berbagai algoritma. Keberhasilan dalam mencapai tujuan dan keseluruhan proses Penambangan Data Pengetahuan (KDD) sangat tergantung pada pemilihan metode atau algoritma yang tepat. Hal ini karena setiap metode memiliki dampak yang signifikan terhadap hasil akhir dan interpretasi data yang dihasilkan. Oleh karena itu, kebijakan yang cermat dalam memilih metode atau algoritma menjadi kunci kesuksesan dalam mengoptimalkan potensi informasi yang dapat ditemukan melalui *data mining* (Naldy & Andri, 2021).

2.5 K-Means

K-Means adalah suatu algoritma pengelompokan yang beroperasi secara iteratif dengan melakukan partisi untuk mengklasifikasikan atau mengelompokkan sejumlah besar objek. Algoritma ini secara berulang melakukan proses pengelompokan dengan membagi objek-objek tersebut ke dalam kluster atau kelompok yang memiliki kesamaan berdasarkan karakteristik tertentu (Triandini et al., 2021). K-Means merupakan salah satu metode *clustering* non-hirarki yang berupaya mempartisi data yang ada ke dalam satu atau lebih kelompok atau kluster. Dalam konteks ini, metode non-hirarki mengacu pada pendekatan di mana kluster tidak memiliki tingkatan atau struktur hirarkis dan objek-objek data dikelompokkan berdasarkan kemiripan mereka ke dalam *cluster* tertentu. Tujuan utama dari K-Means adalah membentuk kelompok-kelompok yang saling homogen dan meminimalkan variasi antara objek-objek dalam satu kelompok dengan objek-objek dalam kelompok lainnya (Triyansyah & Fitrianah, 2018).
Tahapan dari proses K-Means meliputi (Virgo et al., 2020):

1) Inputkan data yang akan dilakukan pengklasteran.
2) Tentukan jumlah kluster yang diinginkan.
3) Tentukan pusat kluster atau *centroid* awal.
4) Lakukan perhitungan jarak Euclidean, proses pengklasteran data, dan perhitungan *centroid* baru untuk iterasi ke-n.
5) Tentukan hasil akhir dari proses pengklasteran.

2.6 *Clustering*

Clustering adalah suatu proses pengelompokan atau pembagian data dalam suatu himpunan menjadi beberapa kelompok, di mana kesamaan data dalam satu kelompok lebih besar dibandingkan dengan kesamaan data tersebut dengan kelompok lainnya. Potensi dari teknik *clustering* ini dapat digunakan untuk mengidentifikasi struktur dalam data, yang nantinya dapat diterapkan dalam berbagai aplikasi seperti klasifikasi, pengolahan gambar, dan pengenalan pola.

Teknik *clustering* memiliki dua metode pengelompokan, yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *Hierarchical clustering* adalah metode pengelompokan data yang mengelompokkan dua data atau lebih yang memiliki kesamaan atau kemiripan. Proses ini dilanjutkan dengan mengelompokkan objek lain yang memiliki kedekatan dua dan proses ini terus berlangsung hingga terbentuk suatu struktur pohon (*tree*) dengan tingkatan hirarki yang jelas antar objek, mulai dari yang paling mirip hingga yang paling tidak mirip. Meskipun secara logika, pada akhirnya, semua objek akan membentuk sebuah *cluster* (Sadewo et al., 2018).

2.7 Elbow *Method*

Metode Elbow adalah teknik yang digunakan dalam analisis data dan pembelajaran mesin untuk menentukan jumlah klaster optimal dalam suatu *dataset*. Metode ini melibatkan plot variasi yang dijelaskan oleh jumlah klaster yang berbeda dan mengidentifikasi titik "Elbow" atau "siku", di mana tingkat variasi menurun tajam dan stabil, menunjukkan jumlah klaster yang tepat untuk analisis atau pelatihan model (Sholeh et al., 2022).

Berikut adalah langkah-langkah metode Elbow dalam klasterisasi K-means (Yudhistira & Andika, 2023):

1) Pilih jumlah klaster untuk *dataset* (K).
2) Pilih k *centroid* secara acak dari *dataset*.
3) Gunakan jarak Euclidean sebagai metrik untuk menghitung jarak titik dari *centroid* terdekat dan menetapkan titik ke *centroid* klaster terdekat, sehingga menciptakan k klaster.
4) Kemudian *centroid* baru dari klaster yang terbentuk.
5) Tetaplam seluruh titik data berdasarkan *centroid* baru ini, lalu ulangi langkah 4.
6) Lanjutkan langkah ini untuk sejumlah iterasi yang diberikan sampai posisi *centroid* tidak berubah, yaitu tidak ada lagi konvergensi.

2.8 RapidMiner

RapidMiner merupakan sebuah platform perangkat lunak ilmu data yang telah dikembangkan oleh perusahaan bernama RapidMiner. Platform ini menyediakan lingkungan terintegrasi yang mencakup berbagai fungsi seperti persiapan data, pembelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis prediktif. Dengan kata lain, RapidMiner menyediakan berbagai alat dan fitur dalam satu kesatuan platform untuk mendukung sejumlah besar kegiatan di bidang ilmu data, mulai dari pengolahan data hingga pengembangan model prediktif. Platform ini dirancang untuk memfasilitasi tugas-tugas analisis data yang kompleks dan memungkinkan para pengguna untuk mengintegrasikan berbagai aspek dari siklus hidup analisis data dalam satu lingkungan yang terpusat (Anjelita et al., 2019).

3. HASIL DAN PEMBAHASAN

3.1 Sumber Data

Sumber data pada penelitian ini yaitu jumlah penduduk dari setiap kota di provinsi Jawa Timur dari tahun 2021-2023 (Badan Pusat Statistik, 2024). Data di dapat dari *website* https://kedirikota.bps.go.id/ yang merupakan sumber data utama karena menyediakan informasi yang akurat dan terverifikasi mengenai statistik kependudukan. Pemilihan situs ini sebagai sumber utama didasarkan pada reputasinya dalam menyajikan data resmi yang dapat diandalkan. Selain itu, penelitian ini juga didukung oleh berbagai literatur terkait dengan metode *clustering* dan penerapan algoritma K-Means. Literatur ini digunakan sebagai acuan teoritis untuk memperkuat metodologi yang digunakan dalam analisis data.

3.2 Menyeleksi Data

Data yang digunakan dalam penelitian ini berasal dari sumber yang dapat dipercaya, yaitu data dari *website* https://kedirikota.bps.go.id/. Data ini mencakup informasi jumlah penduduk di setiap kota di Provinsi Jawa Timur untuk periode tahun 2021 hingga 2023. Pemilihan data yang valid

dan representatif menjadi kunci dalam memastikan keakuratan analisis klasterisasi. Data yang tidak lengkap atau tidak akurat dapat menghasilkan hasil yang bias atau tidak dapat diandalkan. Setelah melalui proses seleksi, *dataset* yang dihasilkan mencakup variabel jumlah penduduk untuk setiap kota, menciptakan data yang valid untuk analisis klasterisasi menggunakan algoritma K-Means. Hasil seleksi data ini akan membentuk landasan yang kuat untuk memahami pola sebaran penduduk di Provinsi Jawa Timur dan membantu mencapai tujuan penelitian yang telah ditetapkan.

3.3 Mengolah Data

Data dari *website* kemudian diolah untuk mendapatkan data yang dapat dianalisis menggunakan algoritma K-Means dengan mudah. Data yang telah diolah akan menjadi *dataset* yang nantinya akan ditentukan klasternya menggunakan algoritma K-Means. Data yang telah di olah terlihat pada Tabel 1.

Tabel 1 *Dataset* Jumlah Penduduk Setiap Kota di Jawa Timur dengan Rentang 2021-2023

Wilayah	2021	2022	2023
Pacitan	589.108	592.916	596.649
Ponorogo	955.839	964.253	972.582
Trenggalek	734.888	739.669	744.358
Tulungagung	1.096.588	1.105.337	1.113.973
Blitar	1.231.013	1.240.322	1.249.497
Kediri	1.644.400	1.656.020	1.667.450
Malang	2.668.296	2.685.900	2.703.175
Lumajang	1.127.094	1.137.227	1.147.261
Jember	2.550.360	2.567.718	2.584.771
Banyuwangi	1.718.462	1.731.731	1.744.814
Bondowoso	778.525	781.417	784.192
Situbondo	688.337	691.260	694.081
Probolinggo	1.155.894	1.159.965	1.163.859
Pasuruan	1.611.805	1.619.035	1.626.029
Sidoarjo	2.091.930	2.103.401	2.114.588
Mojokerto	1.125.522	1.133.584	1.141.516
Jombang	1.325.914	1.335.972	1.345.886
Nganjuk	1.109.683	1.117.033	1.124.247
Madiun	750.143	757.665	765.135
Magetan	674.133	678.343	682.466
Ngawi	873.346	877.432	881.393
Bojonegoro	1.307.602	1.315.125	1.322.474
Tuban	1.203.127	1.209.543	1.215.795
Lamongan	1.356.027	1.371.509	1.386.941
Gresik	1.320.570	1.332.664	1.344.648
Bangkalan	1.071.712	1.086.620	1.101.556
Sampang	976.020	984.162	992.210
Pamekasan	853.507	857.818	862.009
Sumenep	1.129.822	1.136.632	1.143.295
Kota Kediri	287.962	289.418	290.836
Kota Blitar	150.371	151.960	153.541
Kota Malang	844.933	846.126	847.182
Kota Probolinggo	241.202	243.200	245.174
Kota Pasuruan	209.528	211.497	213.450
Kota Mojokerto	133.272	134.350	135.414
Kota Madiun	196.917	199.192	201.460
Kota Surabaya	2.880.284	2.887.223	2.893.698
Kota Batu	214.653	216.735	218.802

3.4 Implementasi K-Means

3.4.1 Menentukan Jumlah k Berdasarkan Elbow *Method*

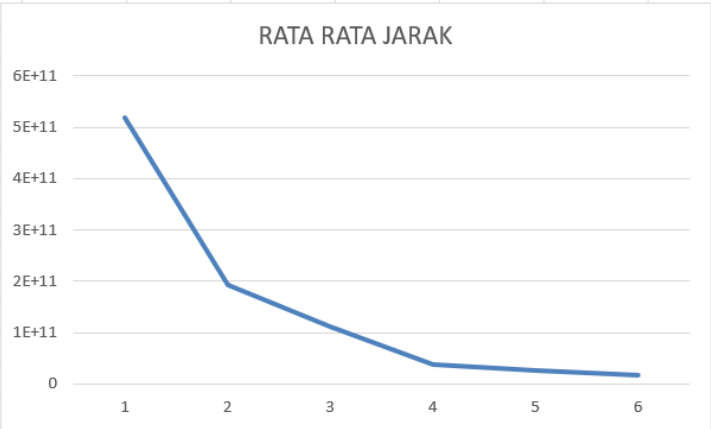
Metode Elbow dipilih dengan cara mengukur performa *distance centroid* menggunakan aplikasi RapidMiner dengan berbagai variasi kluster. Metode Elbow dipilih karena merupakan salah satu metode yang umum digunakan dan relatif mudah dipahami dalam menentukan jumlah klaster yang optimal. Metode Elbow digunakan untuk mengidentifikasi titik di mana penurunan varian antara jumlah klaster yang berbeda menjadi kurang signifikan. Keunggulan metode Elbow adalah kemampuannya untuk memberikan panduan yang jelas dalam menentukan jumlah klaster yang sesuai dengan *dataset* tertentu tanpa memerlukan asumsi sebelumnya tentang distribusi data.

Langkah pertama penentuan jumlah kluster optimal menggunakan metode Elbow adalah berbagai jumlah klaster dipilih dan algoritma K-Means diterapkan untuk masing-masing jumlah klaster tersebut. Kemudian, variasi yang dijelaskan oleh model terhadap jumlah klaster dievaluasi. Grafik yang menunjukkan hubungan antara jumlah klaster dan variasi ini digunakan untuk mengidentifikasi titik "Elbow" atau "siku", yaitu titik di mana penurunan variasi mulai menurun secara signifikan. Pada penelitian ini, titik "siku" terletak pada jumlah klaster tertentu, yang kemudian dipilih sebagai jumlah klaster optimal.

Hasil dari analisis banyak kluster dengan rata-rata jarak dapat ditemukan pada Tabel 2. Berdasarkan grafik metode Elbow pada Gambar 2, terlihat bahwa garis siku terletak di angka 2 pada saat k berjumlah 3 dikarenakan k dimulai dari 0. Sehingga dapat ditentukan jumlah klaster yaitu sebanyak 3 klaster dengan kategori klaster C0, C1, dan C2 dengan kategori masing-masing *cluster* dapat dilihat pada Tabel 3.

Tabel 2 Perhitungan Jarak *Cluster* Setiap Banyak *Cluster*

Banyak *Cluster*	Rata-Rata Jarak
2	5.17948E+11
3	1.93276E+11
4	1.12487E+11
5	38470766310
6	27295170984
7	16188264246



Gambar 2 Grafik Elbow *Method*

Tabel 3 Kategori Klaster

C0	Kota dengan sebaran penduduk sedang
C1	Kota dengan sebaran penduduk terbesar
C2	Kota dengan sebaran penduduk terkecil

3.4.2 Menentukan Pusat Klaster

Berdasarkan data jumlah penduduk dari Tabel 1, maka dapat diambil tiga contoh data sebagai pusat klaster atau pusat *centroid* di mana data pada C0 diambil dari banyak penduduk pada Tuban, data pada C1 diambil dari banyak penduduk dari Jember, dan C2 diambil dari banyak penduduk dari kota Kediri. Contoh data pusat klaster dapat dilihat pada Tabel 4. Proses penentuan pusat klaster ini dilakukan untuk memahami karakteristik masing-masing klaster dan dapat memberikan tinjauan yang lebih jelas terhadap pola sebaran penduduk di Provinsi Jawa Timur.

Tabel 4 Pusat Klaster/*Centroid*

Centroid	2021	2022	2023
C0	1.203.127	1.209.543	1.215.795
C1	2.550.360	2.567.718	2.584.771
C2	287.962	289.418	290.836

3.4.3 Menentukan Euclidean *Distance*

$$D_{(i,j)} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

Dalam menghitung Euclidean *distance* seperti pada Pers. (1), jarak antara data ke-i dan pusat *cluster* ke-j dilambangkan sebagai $D(i,j)$. Jarak ini dihitung untuk menentukan seberapa dekat data ke-i berada dengan pusat *cluster* ke-j. Selain itu, x_{ki} merujuk pada nilai data ke-i pada atribut data ke-k. Ini berarti bahwa setiap data memiliki beberapa atribut atau fitur, dan x_{ki} menunjukkan nilai spesifik dari atribut ke-k untuk data ke-i tersebut. Selanjutnya, x_{kj} digunakan untuk menunjukkan titik pusat *cluster* ke-j pada atribut ke-k. Titik pusat ini adalah nilai rata-rata dari semua data dalam *cluster* tersebut untuk atribut ke-k, yang digunakan sebagai referensi untuk menghitung jarak antara data dan pusat *cluster*. Selanjutnya menentukan Euclidean *distance* berdasarkan *dataset* dengan *centroid* yang ditentukan.

$$D\,(1,1) = \sqrt{(589108 - 1203127)^2} + (592916 - 1209543)^2 + (596649 - 1215795)^2 = 1067984$$

$$D\,(1,2) = \sqrt{(589108 - 2550360)^2} + (592916 - 2567718)^2 + (596649 - 2584771)^2 = 3420377$$

$$D\,(1,3) = \sqrt{(589108 - 287962)^2} + (592916 - 289418)^2 + (596649 - 290836)^2 = 525662$$

Dalam proses pengambilan sampel yang melibatkan tiga langkah, perhitungan jarak antara data ke-1 dan ke-3 terhadap *centroid* data dapat dilakukan dengan menggunakan fungsi SQRT yang tersedia dalam perangkat lunak Microsoft Excel. Selanjutnya, klaster untuk setiap data dapat ditentukan berdasarkan tiga klaster dengan jarak terdekat dari setiap data, atau klaster mana yang memberikan nilai jarak Euclidean terkecil untuk setiap data. Informasi tentang penentuan klaster berdasarkan nilai jarak Euclidean dapat ditemukan dalam Tabel 5.

3.4.4 Menentukan *Centroid* Baru

Dalam tahap menentukan *centroid* Baru, setelah klaster untuk setiap data telah ditentukan, langkah berikutnya adalah menghitung nilai rata-rata dari setiap kolom data pada klaster yang sama. Proses ini dilakukan dengan cara menjumlahkan semua data pada kolom yang sesuai dalam setiap klaster, kemudian hasil penjumlahan tersebut dibagi oleh jumlah data pada kolom tersebut. Misalnya, untuk klaster C0, jumlah semua nilai atribut dari setiap data dalam klaster tersebut dijumlahkan, kemudian hasil penjumlahan tersebut dibagi dengan jumlah data dalam klaster C0. Proses ini bertujuan untuk menemukan pusat klaster yang merupakan representasi rata-rata dari seluruh data dalam klaster tersebut. *Centroid* baru (*C*) dihitung dengan mempertimbangkan jumlah data (*x*) dan banyaknya data (*y*) dalam kolom tertentu (*b*) pada kluster

tertentu (*a*). Proses ini membantu dalam menentukan pusat dari setiap kluster berdasarkan distribusi data yang ada. Maka berdasarkan Pers. (2) terbentuk *centroid* baru pada Tabel 6.

$$C_{ab} = \frac{y}{x} \quad (2)$$

Tabel 5 Penentuan *Cluster* dari Nilai Jarak Euclidean Terkecil pada Iterasi-1

Wilayah	2021	2022	2023	*Cluster*
Pacitan	1.067.984	3.420.377	525.663	C2
Ponorogo	424.819	2.777.184	1.168.863	C0
Trenggalek	813.807	3.166.200	779.840	C2
Tulungagung	180.491	2.532.836	1.413.207	C0
Blitar	53.486	2.299.046	1.646.996	C0
Lumajang	125.330	2.477.585	1.468.464	C0
Bondowoso	741.525	3.093.922	852.124	C0
Situbondo	897.670	3.250.067	695.977	C2
Probolinggo	85.943	2.438.298	1.507.756	C0
Mojokerto	131.566	2.483.926	1.462.115	C0
Jombang	219.054	2.133.378	1.812.664	C0
Nganjuk	160.221	2.512.602	1.433.438	C0
Madiun	782.613	3.134.988	811.065	C0
Magetan	920.026	3.272.420	673.621	C2
Ngawi	575.220	2.927.617	1.018.426	C0
Bojonegoro	182.874	2.169.523	1.776.518	C0
Tuban	0	2.352.397	1.593.646	C0
Lamongan	280.896	2.071.749	1.874.335	C0
Gresik	213.436	2.139.076	1.806.976	C0
Bangkalan	213.144	2.565.147	1.380.964	C0
Sampang	390.339	2.742.708	1.203.336	C0
Pamekasan	609.187	2.961.584	984.458	C0
Sumenep	126.277	2.478.667	1.467.374	C0
Kota Kediri	1.593.646	3.946.040	0	C2
Kota Blitar	1.831.710	4.184.102	238.067	C2
Kota Malang	629.484	2.981.875	964.189	C0
Probolinggo	1.673.686	4.026.079	80.048	C2
Kota Pasuruan	1.728.592	4.080.985	134.952	C2
Kota Mojokerto	1.862.217	4.214.611	268.572	C2
Kota Madiun	1.749.898	4.102.290	156.263	C2
Kota Batu	1.719.518	4.071.911	125.881	C2
Kediri	773.341	1.579.061	2.366.979	C0
Malang	2.557.078	204.684	4.150.722	C1
Jember	2.352.397	0	3.946.040	C1
Banyuwangi	904.495	1.447.914	2.498.129	C0
Pasuruan	709.220	1.643.189	2.302.864	C0
Sidoarjo	1.548.154	804.251	3.141.800	C1
Kota Surabaya	1.067.984	3.420.377	525.663	C1

Tabel 6 *Centroid* Baru

Centroid	2021	2022	2023
C0	1.155.111	1.163.356	1.171.476
C1	2.547.718	2.561.061	2.574.058
C2	374.579	377.140	379.657

Langkah 3.4.3 dan 3.4.4 dilakukan terus menerus sampai tidak ada lagi pergeseran nilai jarak dan pusat klaster serta data klaster. Hal ini dilakukan untuk memastikan bahwa klaster yang

terbentuk sudah stabil dan tidak berubah lagi. Pergeseran data merujuk pada perubahan posisi data dalam klaster. Pada awalnya, data akan dikelompokkan ke dalam klaster yang memiliki jarak terdekat dengan data tersebut. Namun, setelah pusat klaster dihitung kembali, beberapa data mungkin akan berpindah ke klaster lain yang memiliki jarak lebih dekat. Pergeseran data ini akan terus terjadi sampai tidak ada lagi perubahan posisi data dalam klaster.

Dalam proses 3.4.3, jarak antara data dengan pusat klaster dihitung menggunakan persamaan jarak Euclidean. Pada langkah 3.4.3 terbentuk Tabel 4 yang mengelompokkan data ke dalam klaster yang memiliki jarak terdekat dengan data tersebut. Sedangkan proses 3.4.4 dilakukan untuk menghitung kembali pusat klaster dengan keanggotaan klaster yang baru. Jika pusat klaster tidak berubah, maka proses klaster dianggap selesai. Namun, jika pusat klaster masih berubah, maka proses 3.4.3, dan 3.4.4 akan diulang kembali sampai tidak ada lagi perubahan dalam klaster.

3.4.5 Hasil Akhir Klaster

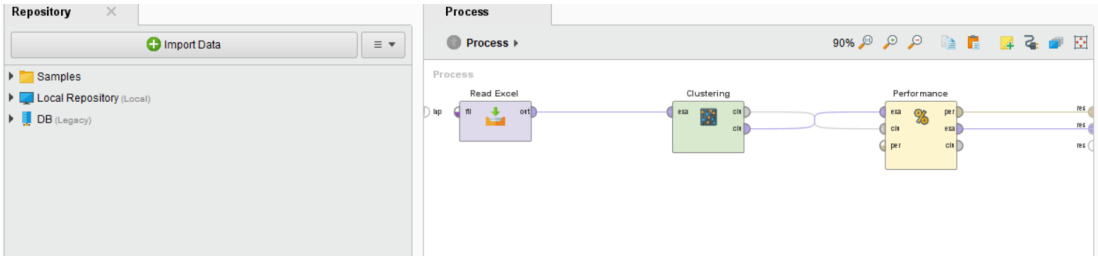
Tabel 7 Hasil Akhir Klaster

Kota	2021	2022	2023	*Cluster*
Ponorogo	955.839	964.253	972.582	C0
Tulungagung	1.096.588	1.105.337	1.113.973	C0
Blitar	1.231.013	1.240.322	1.249.497	C0
Kediri	1.644.400	1.656.020	1.667.450	C0
Lumajang	1.127.094	1.137.227	1.147.261	C0
Banyuwangi	1.718.462	1.731.731	1.744.814	C0
Probolinggo	1.155.894	1.159.965	1.163.859	C0
Pasuruan	1.611.805	1.619.035	1.626.029	C0
Mojokerto	1.125.522	1.133.584	1.141.516	C0
Jombang	1.325.914	1.335.972	1.345.886	C0
Nganjuk	1.109.683	1.117.033	1.124.247	C0
Ngawi	873.346	877.432	881.393	C0
Bojonegoro	1.307.602	1.315.125	1.322.474	C0
Tuban	1.203.127	1.209.543	1.215.795	C0
Lamongan	1.356.027	1.371.509	1.386.941	C0
Gresik	1.320.570	1.332.664	1.344.648	C0
Bangkalan	1.071.712	1.086.620	1.101.556	C0
Sampang	976.020	984.162	992.210	C0
Pamekasan	853.507	857.818	862.009	C0
Sumenep	1.129.822	1.136.632	1.143.295	C0
Kota Malang	844.933	846.126	847.182	C0
Malang	2.668.296	2.685.900	2.703.175	C1
Jember	2.550.360	2.567.718	2.584.771	C1
Sidoarjo	2.091.930	2.103.401	2.114.588	C1
Kota Surabaya	2.880.284	2.887.223	2.893.698	C1
Pacitan	589.108	592.916	596.649	C2
Trenggalek	734.888	739.669	744.358	C2
Bondowoso	778.525	781.417	784.192	C2
Situbondo	688.337	691.260	694.081	C2
Madiun	750.143	757.665	765.135	C2
Magetan	674.133	678.343	682.466	C2
Kota Kediri	287.962	289.418	290.836	C2
Kota Blitar	150.371	151.960	153.541	C2
Kota Probolinggo	241.202	243.200	245.174	C2
Kota Pasuruan	209.528	211.497	213.450	C2
Kota Mojokerto	133.272	134.350	135.414	C2
Kota Madiun	196.917	199.192	201.460	C2
Kota Batu	214.653	216.735	218.802	C2

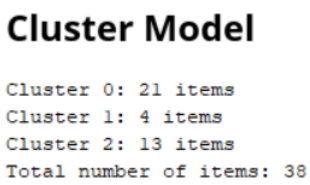
Setelah menyelesaikan proses pada tahap 3.4.3 dan 3.4.4, tidak ditemukan adanya perubahan posisi data dalam klaster. Hal ini menunjukkan bahwa algoritma K-Means telah mencapai konvergensi, di mana posisi setiap data dalam klaster sudah stabil dan tidak lagi berpindah. Oleh karena itu, data yang diperoleh pada Tabel 7 dapat dianggap sebagai hasil akhir dari proses klasterisasi yang telah dilakukan.

3.5 Implementasi K-Means Menggunakan RapidMiner

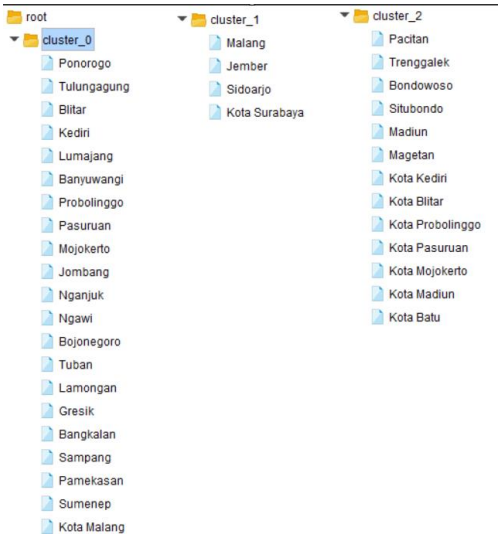
Dalam implementasi algoritma K-Means menggunakan perangkat lunak RapidMiner, *dataset* pada Tabel 1 dijadikan sebagai data yang akan dianalisis yang tersedia dalam format file Excel. Data tersebut kemudian diimpor ke dalam RapidMiner untuk memulai proses analisis. Langkah-langkah analisis yang dilakukan dapat divisualisasikan pada Gambar 3, yang memberikan gambaran jelas dan memudahkan pemahaman tentang proses analisis yang dilakukan.



Gambar 3 Tampilan Input *Dataset* dan Pengujian Menggunakan Performance



Gambar 4 Jumlah Klaster



Gambar 5 Sebaran Klaster

Dari Gambar 3, terlihat bahwa menu performance, yang ditunjukkan oleh kotak berwarna kuning, digunakan untuk membuat grafik metode Elbow dari setiap nilai k yang digunakan. Grafik Elbow *method* ini menunjukkan bagaimana jumlah perbedaan antara titik data dan pusat klaster berubah

seiring dengan penambahan jumlah klaster. Tujuan dari grafik ini adalah untuk menemukan titik di mana penurunan perbedaan antara titik data dan pusat klaster mulai melambat secara signifikan, membentuk titik yang menyerupai siku atau "*elbow*", yang menandakan jumlah klaster yang optimal.

Hasil dari perhitungan menggunakan RapidMiner dan perhitungan manual menunjukkan bahwa kedua metode tersebut menghasilkan data yang sama baik dalam jumlah klaster maupun posisi klaster yang dihasilkan. Informasi mengenai jumlah data dalam setiap klaster dan total keseluruhan data dapat ditemukan pada Gambar 4.

Sebaran data pada setiap *cluster* dapat dilihat pada Gambar 5, di mana tergambar dengan jelas sebaran data pada masing-masing *cluster* yang dihasilkan oleh algoritma K-Means. Setiap *cluster* menampung sejumlah nama kota yang sesuai dengan karakteristiknya, mulai dari *cluster* 0, *cluster* 1, hingga *cluster* 2. Hal ini mengindikasikan bahwa algoritma K-Means telah berhasil mengelompokkan kota-kota dalam *dataset* ke dalam tiga kelompok yang berbeda berdasarkan atribut-atribut yang dimiliki.

4. KESIMPULAN

Dalam pengolahan data jumlah penduduk di provinsi jawa timur menggunakan algoritma k-means melalui RapidMiner dan perhitungan manual, ditemukan kesimpulan bahwa kedua metode tersebut menghasilkan jumlah klaster yang serupa yaitu sebanyak 3 klaster dengan *Cluster* 0, yaitu kota dengan sebaran penduduk sedang, mencakup 21 kota, *Cluster* 1, yaitu kota dengan sebaran penduduk terbesar, mencakup 4 kota dan *Cluster* 2, yaitu kota dengan sebaran penduduk terkecil, mencakup 13 kota. Hasil analisis klasterisasi memberikan gambaran yang konsisten terkait pola sebaran penduduk di setiap kota. Kesamaan ini memberikan validitas terhadap aplikasi algoritma k-means dalam konteks analisis demografis provinsi jawa timur.

Oleh karena itu, hasil penelitian ini memberikan keyakinan bahwa algoritma k-means dengan metode Elbow dapat digunakan secara efektif untuk mengklasterisasi data jumlah penduduk, dengan hasil yang konsisten baik melalui pendekatan RapidMiner maupun perhitungan manual. Hasil temuan ini tidak hanya memberikan gambaran yang lebih jelas tentang pola persebaran penduduk di Jawa Timur, tetapi juga berkontribusi memudahkan pemerintah untuk merancang kebijakan yang lebih tepat sasaran berdasarkan pola sebaran dari setiap kota. Ini akan membantu dalam mengalokasi sumber daya yang lebih efisien, perencanaan tata ruang kota yang berkelanjutan, pengelolaan resiko bencana, serta merencanakan pembangunan infrastruktur yang sesuai dengan kebutuhan masing-masing kota sehingga kebijakan-kebijakan yang dilakukan lebih tepat sasaran pada setiap kota di Provinsi Jawa Timur.

DAFTAR PUSTAKA

Anjelita, M., Windarto, A. P., & Hartama, D. (2019). Pemanfaatan Data Mining pada Pengelompokan Provinsi Terhadap Pencemaran Lingkungan Hidup. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1). https://doi.org/10.30865/komik.v3i1.1675

Anufia, B., & Alhamid, T. (2019). *Instrumen Pengumpulan Data*. OSF. https://doi.org/10.31227/OSF.IO/S3KR6

Badan Pusat Statistik. (2024, March 14). *Jumlah Penduduk Provinsi Jawa Timur (Jiwa), 2021-2023*. Badan Pusat Statistik Kota Kediri. https://kedirikota.bps.go.id/indicator/12/358/1/jumlah-penduduk-provinsi-jawa-timur.html

Budiana, N. D., Siregar, R. R. A., & Susanti, M. N. I. (2019). Penetapan Instruktur Diklat Menggunakan Metode Clustering K-Means dan Topsis Pada PT PLN (Persero) Udiklat Jakarta. *PETIR*, *12*(2), 111–121. https://doi.org/10.33322/petir.v12i2.454

Dewi, F. P., Aryni, P. S., & Umaidah, Y. (2022). Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(2), 111–121. https://doi.org/10.14421/jiska.2022.7.2.111-121

Fahrozi, A. Al, Insani, F., Budianita, E., & Afrianty, I. (2023). Implementasi Algoritma K-Means dalam Menentukan Clustering pada Penilaian Kepuasan Pelanggan di Badan Pelatihan Kesehatan Pekanbaru. *Indonesian Journal of Innovation Multidisipliner Research*, *1*(4), 474–492. https://doi.org/10.31004/IJIM.V1I4.53

Fitriyah, H., Safitri, E. M., Muna, N., Khasanah, M., Aprilia, D. A., & Nurdiansyah, D. (2023). Implementasi Algoritma Clustering dengan Modifikasi Metode Elbow untuk Mendukung Strategi Pemerataan Bantuan Sosial di Kabupaten Bojonegoro. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *4*(3), 1598–1607. https://doi.org/10.46306/lb.v4i3.453

Harahap, L. M., Fuadi, W., Rosnita, L., Darnila, E., & Meiyanti, R. (2022). Klastering Sayuran Unggulan Menggunakan Algoritma K-Means. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(3), 567–579. https://doi.org/10.28932/jutisi.v8i3.5277

Herviany, M., Putri Delima, S., Nurhidayah, T., & Kasini, K. (2021). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokkan Daerah Rawan Tanah Longsor Pada Provinsi Jawa Barat. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 34–40. https://doi.org/10.57152/malcom.v1i1.60

Naldy, E. T., & Andri, A. (2021). Penerapan Data Mining Untuk Analisis Daftar Pembelian Konsumen Dengan Menggunakan Algoritma Apriori Pada Transaksi Penjualan Toko Bangunan MDN. *Jurnal Nasional Ilmu Komputer*, *2*(2), 89–101. https://doi.org/10.47747/jurnalnik.v2i2.525

Nawassyarif, M. Julkarnain, & Rizki Ananda, K. (2020). Sistem Informasi Pengolahan Data Ternak Unit Pelaksana Teknis Produksi dan Kesehatan Hewan Berbasis Web. *Jurnal Informatika, Teknologi Dan Sains*, *2*(1), 32–39. https://doi.org/10.51401/jinteks.v2i1.556

Nofiar, A., Defit, S., & Sumijan. (2019). Penentuan Mutu Kelapa Sawit Menggunakan Metode K-Means Clustering. *Jurnal KomtekInfo*, *5*(3), 1–9. https://doi.org/10.35134/komtekinfo.v5i3.26

Sadewo, M. G., Windarto, A. P., & Wanto, A. (2018). Penerapan Algoritma Clustering dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/Mitigasi Bencana Alam Menurut Provinsi dengan K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 311–319. https://doi.org/10.30865/komik.v2i1.943

Sholeh, M., Suraya, S., & Andayati, D. (2022). Machine Linear untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *8*(1), 20–27. https://doi.org/10.26418/JP.V8I1.48822

Syahfitri, N., Budianita, E., Nazir, A., & Afrianty, I. (2023). Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(3), 1668–1675. https://doi.org/10.30865/KLIK.V4I3.1525

Talakua, M. W., Leleury, Z. A., & Taluta, A. W. (2017). Analisis Cluster dengan Menggunakan Metode K-Means untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *11*(2), 119–128. https://doi.org/10.30598/barekengvol11iss2pp119-128

Triandini, M., Defit, S., & Nurcahyo, G. W. (2021). Data Mining dalam Mengukur Tingkat Keaktifan Siswa dalam Mengikuti Proses Belajar pada SMP IT Andalas Cendekia. *Jurnal Informasi Dan Teknologi*, 167–173. https://doi.org/10.37034/jidt.v3i3.120

Triyansyah, D., & Fitrianah, D. (2018). Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing. *Jurnal Telekomunikasi Dan Komputer*, *8*(3), 163–182. https://doi.org/10.22441/incomtech.v8i3.4174

Virgo, I., Defit, S., & Yuhandri, Y. (2020). Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering. *Jurnal Sistim Informasi Dan Teknologi*, *2*(1), 23–28. https://doi.org/10.37034/jsisfotek.v2i1.17

Wicaksana, R. S., Heksaputra, D., Syah, T. A., & Nur'aini, F. F. (2023). Pendekatan K-Means Clustering Metode Elbow Pada Analisis Motivasi Pengunjung Festival Halal JHF#2. *Jurnal Ilmiah Ekonomi Islam*, *9*(3), 4162. https://doi.org/10.29040/jiei.v9i3.10591

Yudhistira, A., & Andika, R. (2023). Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering. *Journal of Artificial Intelligence and Technology Information (JAITI)*, *1*(1), 20–28. https://doi.org/10.58602/jaiti.v1i1.22