ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 13, Nomor 2, Tahun 2024, Halaman 351 - 362
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS KLASTER KECAMATAN DI KOTA SURABAYA BERDASARKAN DATA PENDIDIKAN TAHUN 2022-2023

Nanda Reza Handitia[1*], A'yunin Sofro[2]
[1,2]Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya
***e-mail**: ayuninsofro@unesa.ac.id

DOI: 10.14710/j.gauss.13.2.351-362

Article Info:
Received: 2024-08-01
Accepted: 2024-11-18
Available Online: 2024-11-21

Keywords:
Education; K-Means; K-Medoids; Fuzzy C-Means

Abstract: Education is an essential requirement for every individual to develop a country's human resources. The national education law emphasizes the importance of equal access, improving quality, and efficient education management to face global challenges. However, the equitable distribution of education in Indonesia, particularly through the zoning system, still faces significant challenges in major cities like Surabaya due to unequal distribution of educational access and facilities. This study compares the effectiveness of three non-hierarchical *clustering* analysis methods on Surabaya City's education data for 2022-2023. The data used was obtained from the latest publication of BPS Surabaya titled "Surabaya Municipality in Figures 2024". The data includes education data from 31 sub-districts in Surabaya, including the number of schools, students, and teachers at each level of education, namely elementary school, junior high school, and senior high school. The results of this research indicate that the *K-Means* method has the highest average coefficient value, with an average *Silhouette Coefficient* of 0.592. Therefore, the *K-Means* method has the most optimal *cluster* accuracy compared to other methods. These findings emphasize the need for more attention to sub-districts with low educational conditions to ensure equal access to education throughout the city of Surabaya.

1. PENDAHULUAN

Pendidikan merupakan kebutuhan fundamental setiap manusia untuk menyiapkan kualitas *human resources* (Kurniawan et al., 2021). Sesuai UU No. 20 Tahun 2003 tentang sistem pendidikan nasional, pemerintah diwajibkan menjamin pemerataan perolehan pendidikan, peningkatan kualitas, relevansi, dan efektivitas manajemen pendidikan untuk mengatasi tantangan terkait desakan peralihan kehidupan dalam tingkatan regional, nasional, dan internasional. Dalam penyelenggaraan pendidikan yang efisien dan efektif dibutuhkan faktor pendukung dalam kegiatan belajar mengajar, yakni berupa penyediaan fasilitas belajar dan tenaga pendidik yang menguasai media pembelajaran (Sari, 2019). Namun, pemerataan pendidikan di Indonesia masih menghadapi kendala, salah satunya penerapan sistem zonasi yang bertujuan menciptakan keadilan akses pendidikan tetapi tidak diimbangi dengan pemerataan fasilitas pendidikan, seperti jumlah sekolah, murid, dan guru di setiap kecamatan.

Surabaya merupakan salah satu kota besar di Indonesia yang memiliki dinamika pendidikan kompleks. Berdasarkan data pendidikan tahun 2021-2023 menunjukkan ketimpangan dalam jumlah sekolah, murid, dan guru antar kecamatan. Beberapa kecamatan memiliki konsentrasi sekolah tinggi namun jumlah murid tidak sebanding, sementara kecamatan lain menghadapi kekurangan fasilitas. Fenomena ini memerlukan analisis lebih lanjut untuk memahami distribusi pendidikan dan dampaknya terhadap sistem zonasi.

Analisis klaster merupakan salah satu metode data mining yang cocok untuk mengidentifikasi pola dalam data pendidikan. Dengan mengelompokkan kecamatan berdasarkan karakteristik jumlah sekolah, murid, dan guru, analisis ini dapat membantu menentukan wilayah dengan kebutuhan pendidikan paling mendesak. Metode non-hierarki seperti K-Means, K-Medoids, dan Fuzzy C-Means dipilih karena keunggulannya masing-masing. K-Means mampu mengelompokkan data besar secara efisien, K-Medoids lebih tahan terhadap outlier, sementara Fuzzy C-Means memungkinkan keanggotaan ganda yang relevan untuk data pendidikan yang kompleks.

Beberapa penelitian sebelumnya terkait perbandingan metode *clustering* non hirarki telah dilakukan, salah satunya terkait perbandingan algoritma *K-Means* dan *K-Medoids* untuk pengelompokan program BPJS Ketenagakerjaan (Meiriza et al., 2023). Ketiga metode tersebut juga pernah diterapkan pada bidang pendidikan seperti penelitian yang bertujuan untuk pengelompokan data guru di Indonesia (Idris et al., 2019) dan pemetaan penyebaran guru di Provinsi Banten (Priambodo & Prasetyo, 2018). Namun, penelitian-penelitian tersebut cenderung berfokus pada kondisi persebaran di tingkat provinsi dan kota/kabupaten, sementara belum banyak penelitian yang fokus pada satu kota. Maka dibutuhkan pengklasteran yang fokus pada kondisi dari setiap kecamatan karena di sana dampak dari sistem zonasi paling terasa. Pada penelitian ini akan dibandingkan hasil *cluster* menggunakan metode *K-Means*, *K-Medoids*, dan *Fuzzy C-Means* dalam menentukan metode mana yang paling efektif untuk pengelompokan data pendidikan di Surabaya. Hasil pengelompokan yang akurat diharapkan dapat menjadi dasar pengambilan kebijakan untuk pemerataan pendidikan yang lebih baik, sehingga setiap kecamatan memiliki akses yang setara terhadap fasilitas pendidikan berkualitas.

2. TINJAUAN PUSTAKA

Kondisi pendidikan saat ini belum mencapai standar yang dimandatkan oleh undang-undang. Pemerataan masih jauh dari kata sempurna (Reay, 2020; Tchamyou, 2020). Ketidakmerataan pendidikan terdiri dari dua komponen, yakni mutu pendidikan (layanan pendidikan) dan kuantitas pendidikan (persebaran sekolah, pengaksesan, dan proporsi total sekolah dengan total penduduk) (Hakim, 2016). Persebaran sekolah menjadi aspek penting dalam menjamin akses pendidikan yang adil, peningkatan jumlah sekolah dan distribusi yang merata diharapkan dapat mengurangi kesenjangan akses pendidikan antar wilayah. Selain itu, distribusi siswa juga menjadi faktor penting dalam dinamika pendidikan, mempengaruhi kapasitas dan keefektifan pembelajaran di setiap sekolah. Guru memegang kunci penting dalam proses pembelajaran di sekolah. Pendistribusian guru yang tidak merata di Indonesia adalah rintangan dalam usaha meningkatkan kualitas pendidikan (Nurfatimah et al., 2022). Melalui Menteri Pendidikan dan Kebudayaan Nadiem Makarim, pemerintah telah berupaya membuat kebijakan baru untuk menyelesaikan persoalan distribusi guru, misalnya dengan program zonasi pendidikan yang menyasar pada pemerataan guru profesional (Haekal, 2022).

Sebelum melakukan analisis, perlu pengecekan terkait asumsi pada analisis klaster. Terdapat dua asumsi dalam analisis klaster, yaitu sampel representatif atau terdapat multikolinearitas. Sampel representatif berarti sampel yang diambil sudah mewakili populasi. Sampel representatif dapat diuji menggunakan uji *Kaiser-Mayer-Olkin (KMO). KMO* sering dimanfaatkan dalam menguji syarat kelayakan dari data. Untuk menghitung nilai KMO dapat digunakan rumus sebagai berikut (Widarjono, 2010):

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1\ k\neq j}^{p} r^2 X_j X_k}{\sum_{j=1}^{p}\sum_{k\neq j}^{p} r^2 X_j X_k + \sum_{j=1}^{p}\sum_{k\neq j}^{p} \rho^2 X_j X_k, X_1} \quad (1)$$

dimana:

p : jumlah atribut
rX_jX_k : hubungan antara atribut X_j dan X_k
$\bar{X}_j$: mean dari atribut X_j
$\bar{X}_k$: mean dari atribut X_k
N : jumlah observasi (objek)
$\rho X_j X_k, X_1$: hubungan secara parsial antara atribut X_j dan X_k dengan X_1

Jika didapatkan nilai KMO berada pada rentang 0,5 hingga 1 maka data dianggap data representatif dan analisis faktor layak dilakukan. Selanjutnya, dilakukan *Bartlett Test of Sphericity,* yakni bertujuan analisis ada tidaknya hubungan antara variabel dalam multivariat agar analisis dapat dilanjutkan (Putri & Fithriasari, 2015). Matriks korelasi dikatakan matriks identitas jika atribut bersifat independen yang menandakan di antara peubah tidak terdapat korelasi. Hipotesis percobaan yang dipakai pada *Bartlett Test* adalah sebagai berikut:

H_0 : Matriks korelasi sama dengan matriks identitas (tidak ada multikolinearitas)
H_1 : Matriks korelasi tidak sama dengan matriks identitas

Nilai statistik *Bartlett Test* dapat ditulis dengan persamaan berikut:

$$X^2_{obs} = -\left[(N-1)\frac{(2p+5)}{6} in|\boldsymbol{R}|\right] \quad (2)$$

dimana:
N : total pengamatan
$|\boldsymbol{R}|$: determinan matriks pengamatan
p : banyaknya atribut

Dalam membuat keputusan *Bartlett Test* jika $X^2_{obs} > X^2_{\alpha,\frac{p(p-1)}{2}}$ atau *pvalue* < α maka bermakna bahwa terjadi multikolinearitas antar atribut.

Setelah melakukan uji *KMO* dan *Bartlett*, proses analisis data dapat dilanjutkan dengan normalisasi. Transformasi data melalui normalisasi bisa diselesaikan menggunakan berbagai metode, yaitu *z-score normalization, decimal scaling, Min-Max normalization, softmax, dan sigmoid*. Normalisasi menggunakan *Min-Max* adalah penggunaan transformasi linier data asli untuk memperoleh nilai perbandingan antara data sebelum dan sesudah proses yang seimbangan. Metode ini menghasilkan proporsi data yang seimbang antara satu dengan yang lain (Tasmalaila Hanifa et al., 2017). Metode normalisasi *Min-Max* dapat menggunakan rumus sebagai berikut:

$$v' = \frac{v - min_A}{max_A - min_A}(range.max - range.min) + range.min \quad (3)$$

v' merupakan nilai dari data sesudah normalisasi, *v* adalah data sebelum normalisasi, *minA* merupakan nilai minimal pada atribut sebelum normalisasi, *maxA* adalah nilai maksimal pada atribut sebelum normalisasi, *range min* adalah 0, dan *range max* adalah 1.

Langkah awal dalam analisis klaster non-hirarki adalah penentuan total klaster, yang mana pada penelitian ini digunakan *elbow method*. *Elbow method* adalah salah satu metode untuk memutuskan total klaster optimal dengan melihat persentase hasil membandingkan total klaster yang membentuk siku pada suatu titik (Madhulata, 2012). Hasil persentase setiap nilai klaster ditampilkan dalam bentuk grafik sebagai sumber informasi (Merliana et al., 2019). Dikatakan nilai klaster tersebut paling baik jika pada grafik membentuk siku atau mengalami penurunan yang paling besar (Bholowalia & Kumar, 2014). Nilai SSE merupakan jumlah rata-rata jarak *euclidean* dari setiap titik ke *centroid*. Untuk mendapatkan nilai SSE digunakan persamaan sebagai berikut (Merliana et al., 2019):

$$SSE = \sum_{k=1}^{K} \sum_{x_i \epsilon S_k} \|\boldsymbol{x_i} - \boldsymbol{c_k}\|^2 \quad (4)$$

dimana:
K : banyak klaster
xi : data ke-i
Sk : himpunan dari elemen klaster *k*
ck : rata-rata (pusat) dalam klaster *k*

K-Means adalah algoritma pendekatan non hierarki dengan melakukan partisi pada data ke dalam klaster sehingga data dengan sifat sejenis digolongkan pada suatu klaster, sebaliknya data dengan sifat berbeda ditempatkan di klaster lainnya. Secara sederhana, proses *K-Means* dibagi menjadi 5 langkah berikut (Maryani et al., 2018):

1. Menentukan nilai total klaster (*k*), dalam penelitian ini disesuaikan dengan hasil *elbow method*.
2. Tentukan nilai pusat (*centroid*) awal dari setiap klaster dengan nilai sembarang. Dalam penentuan centroid awal, nilai-nilai centroid dapat diinisialisasi secara sembarang dalam rentang data yang ada, misalnya antara nilai minimum dan maksimum dari setiap atribut.
3. Mengalokasikan seluruh objek pada klaster dengan jarak terdekat terhadap tiap *centroid* menggunakan teori jarak euclidean menggunakan rumus sebagai berikut:

$$D(a,b) = \sqrt{\sum_{k=1}^{n}(a_k - b_k)^2} \quad (5)$$

dimana:
n : banyak atribut
$a_k - b_k$: data dengan atribut *k*

4. Perbarui nilai *centroid* dengan menghitung rata-rata dari semua objek dalam klaster tersebut. Hal ini bertujuan agar centroid lebih representatif terhadap klaster masing-masing.
5. Perhitungan ulang jarak dan pengelompokan dilakukan secara iteratif, dengan kembali ke langkah (3) hingga (4) sampai tidak terjadi perubahan posisi centroid atau pengelompokan objek. Algoritma berhenti ketika tidak ada perubahan lebih lanjut, yang menunjukkan bahwa seluruh objek telah dialokasikan ke klaster secara stabil.

K-Medoids memakai *medoid* untuk pusat klaster, yaitu menggunakan salah satu data yang ada dengan inisialisasi secara acak untuk mewakili klaster (Meiriza et al., 2023). Langkah-langkah algoritma *K-Medoids* dibagi menjadi 6 langkah berikut:

1. Menentukan nilai klaster (*k*), dalam penelitian ini disesuaikan dengan hasil *elbow method*.
2. Memilih sembarang objek pada setiap klaster untuk menjadi medoids. Penentuan medoids dilakukan secara acak berarti bisa dipilih sebarang titik yang.
3. Mengalokasikan semua objek pada klaster dengan jarak terdekat ke masing-masing medoid memakai teori jarak euclidean dan menggunakan rumus seperti persaman 5.
4. Hitung total cost (TC). Total Cost diperoleh dengan menjumlahkan jarak minimum yang terpilih pada semua klaster, disimbolkan d_s. Secara umum TC dapat dihitung menggunakan persamaan berikut:

$$TC = \sum_{i=1}^{N} d_s \quad (6)$$

5. Pilih secara acak kandidat medoids baru seperti pada langkah (2). Kemudian lakukan langkah 3 dan langkah 4. Total Cost terbaru disimbolkan dengan TC_n.
6. Menentukan total simpangan (S) dengan cara menghitung nilai total jarak baru dikurangi total jarak sebelumnya, yaitu S= TC_n. – TC_0. Jika $S > 0$, maka iterasi

berhenti dan anggota klaster yang digunakan adalah anggota saat menghitung TC_0. Apabila S < 0, maka harus menukar objek dengan n data klaster sehingga terbentuk sekumpulan k objek dan dilakukan perulangan untuk langkah (3) hingga (5).

Metode *Fuzzy C-Means* adalah algoritma untuk mengelompokkan data berdasarkan derajat keanggotaan. Selama proses perhitungan, nilai derajat keanggotaan setiap objek akan selalu dihitung dan digunakan untuk menetapkan objek ke dalam satu atau lebih klaster. (Nithila & Kumar, 2016). Data dikelompokkan berdasarkan derajat keanggotaannya, yakni dalam rentang 0 hingga 1, dan ada beberapa jenis data hanya menampilkan keanggotaan sebagian. Algoritma *Fuzzy C-Means* bisa menghasilkan pengelompokan untuk objek tidak teratur dan tersebar. Secara sederhana, proses algoritma *Fuzzy C-Means* dibagi menjadi 7 langkah (Andriyani et al., 2013) sebagai berikut:

1. Memasukkan data yang akan dikelompokkan berupa matriks **X** berukuran n x m (n = banyaknya sampel data dan m = banyaknya atribut setiap data). Xij = data sampel ke-i (i = 1, 2, ..., n), atribut ke-j (j = 1, 2, ..., m)
2. Menentukan nilai-nilai yang diperlukan sebelum melakukan perhitungan, yakni sebagai berikut:
   - Jumlah klaster yang akan dibentuk (c) = $2 \leq c < (n)$
   - Pangkat pembobot (w) = $w \geq 1$; w=2 sering digunakan untuk pembobot karena dianggap paling halus.
   - Jumlah maksimum pengulangan iterasi = *MaxIter*.
   - Jumlah error terkecil yang diharapkan (ξ) = $\xi > 0$
   - Fungsi obyektif awal ($P_0 = 0$)
   - Iterasi awal (t = 1)
3. Membangkitkan matriks partisi awal Unxc = [µik], µik merupakan bilangan random yang menunjukkan derajat keanggotaan, di mana $i = 1,2, \ldots n$
4. Menghitung pusat klaster ke-*k* (*Vkj*) menggunakan *k* = 1, 2, ..., *c* dan *j*=1, 2, ..., *m* dengan rumus persamaan 7:
$$V_{kj} = \frac{\sum_{i=1}^{n}(\mu_{ik})^w . X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^w} \quad (7)$$
5. Mencari nilai fungsi objektif untuk iterasi ke-*t*, dimana *Pt* menyatakan jumlah jarak data ke pusat klaster yang dapat dihitung dengan rumus persamaan 8 berikut:
$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w \quad (8)$$
dimana:
P_t : fungsi objektif
X_{ij} : elemen *X* baris *i*, kolom *j*
V_{kj} : pusat klaster
6. Mencari nilai perubahan matriks partisi melalui rumus persamaan 9 berikut:
$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij}-V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij}-V_{kj})^2\right]^{\frac{-1}{w-1}}} \quad (9)$$
dimana
i : 1, 2, ..., n
k : 1, 2, ..., c
X_{ij} : sampel data ke-*i*, atribut ke-*j*
V_{kj} : pusat klaster ke-*k* untuk atribut ke-*j*
w : pangkat pembobotan
7. Melakukan pemeriksaan kondisi untuk tetap melakukan iterasi atau stop, dengan pemutusan keputusan sebagai berikut:

- Jika ($|Pt - Pt{-}1| < \xi$) atau ($t > MaxIter$) maka stop
- Jika tidak maka t = t + 1, lalu dilakukan perulangan untuk langkah (4) hingga (7).

Setelah klaster terbentuk maka perlu dilakukan validasi hasil klaster untuk menentukan metode yang memiliki nilai akurasi paling besar. *Silhouette Coefficient* berfungsi dalam menentukan seberapa bagus objek diletakkan dalam suatu klaster setelah proses *clustering*, dengan kata lain berupa nilai kualitas dan kekuatan dari klaster. Metode *Silhouette Coefficient* adalah perpaduan dari 2 metode, yakni cohesion dan separation (Dewi & Pramita, 2019). Langkah-langkah dalam perhitungan *Silhouette Coefficient* dibagi menjadi 3 langkah sebagai berikut (Mario et al., 2016):

1. Menghitung rata-rata jarak dari dokumen, misalkan *i* dengan semua dokumen berada dalam satu klaster, digunakan rumus persamaan 10 berikut:

$$a(i) = \frac{1}{|\boldsymbol{A}|-1} \sum j \in_{A, j \neq i} d(i,j) \tag{10}$$

dimana *j* merupakan dokumen lain dalam satu klaster *A* dan *d(i,j)* adalah jarak antara dokumen *i* dengan *j*

2. Menghitung rata-rata jarak dari dokumen *i* tersebut dengan semua dokumen di klaster lain, kemudian diambil nilai terkecilnya melalui persamaan 11 berikut:

$$d(i,C) = \frac{1}{|\boldsymbol{A}|} \sum j \in C \; d(i,j) \tag{11}$$

dimana *d(i,C)* adalah jarak rata-rata dokumen i dengan seluruh objek pada klaster lain yang mana $\boldsymbol{A} \neq C$, lalu $b(i) = \min C \; \neq \boldsymbol{A} \; d(i,j)$.

3. Menentukan nilai koefisien *silhouette* dengan rumus persamaan 12 berikut:

$$s(i) = \frac{b(i) - a(i)}{\max (a(i), b(i))} \tag{12}$$

Berikut ini adalah kriteria pengukuran nilai *Silhouette Coefficient* disajikan pada Tabel 1 sebagai berikut:

Tabel 1. Kriteria *Silhouette Coefficient*

Koefisien *Silhouette*	**Kriteria Klaster**
0,71 - 1,00	*Strong*
0,51 - 0,70	*Good*
0,26 - 0,50	*Weak*
0,00 - 0,25	*Bad*

Jika dalam proses *clustering* dilakukan perbandingan dari beberapa metode, maka analisis karakter dari klaster cukup dilakukan pada hasil pengklasteran dengan metode terbaik (Nahdliyah et al., 2019). Setelah klaster dengan metode terbaik terbentuk maka langkah berikutnya adalah memberi karakteristik yang bisa menggambarkan isi klaster tersebut menurut kesamaan antar objek yang diteliti (Ls et al., 2021). Karakteristik pada setiap klaster yang terbentuk bisa diwakilkan dengan nilai mean dari setiap atribut (Nahdliyah et al., 2019).

3. METODE PENELITIAN

Pada penelitian ini jenis data yang digunakan adalah data sekunder yang diperoleh dari publikasi terbaru Badan Pusat Statistik (BPS) Kota Surabaya berjudul "Kota Surabaya dalam Angka 2024". Data yang diambil mencakup informasi data pendidikan pada tahun 2022-2023 dari 31 kecamatan di Kota Surabaya.

Dalam penelitian ini data yang diambil mencakup informasi data pendidikan pada tahun 2022-2023 dari 31 kecamatan di Kota Surabaya terkait jumlah sekolah, jumlah murid, dan jumlah guru pada masing-masing jenjang pendidikan, yaitu jumlah sekolah SD/MI (X_1), jumlah murid SD/MI (X_2), jumlah guru SD/MI (X_3), jumlah sekolah SMP/MTS (X_4), jumlah

murid SMP/MTS (X_5), jumlah guru SMP/MTS (X_6), jumlah sekolah SMA/SMK/MA (X_7), jumlah murid SMA/SMK/MA (X_8), dan jumlah guru SMA/SMK/MA (X_9). Proses analisis data yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Pengumpulan data pendidikan pada tahun 2022-2023 dari 31 kecamatan di Kota Surabaya.
2. Uji kelayakan data menggunakan metode *Bartlett Test of Sphericity* dan *Kaiser-Mayer-Olkin* (KMO).
3. Normalisasi data.
4. Menentukan jumlah klaster mengunakan data hasil normalisasi menggunakan metode *elbow.*
5. Analisis klaster menggunakan metode K-Means:
   a) Memilih *centroid* awal secara acak misalnya menggunakan data Kecamatan untuk menyatakan nilai setiap klaster.
   b) Menghitung jarak antara setiap data dengan masing-masing *centroid* menggunakan rumus jarak *euclidan.*
   c) Menentukan setiap Kecamatan masuk ke dalam klaster yang mana berdasarkan jarak paling dekat dengan *centroid.*
   d) Memperbarui nilai *centroid* untuk setiap klaster dengan menghitung rata- rata data yang ada pada klaster yang sama.
   e) Dilakukan perulangan langkah (c) dan (d) hingga tidak terjadi perubahan pada *centroid* dan keseluruhan objek tidak dapat diklasifikasikan lagi
6. Analisis klaster menggunakan metode K-Medoids:
   a) Memilih *medoid* awal secara acak misalnya menggunakan data Kecamatan untuk menyatakan nilai setiap klaster.
   b) Menghitung jarak antara setiap data dengan masing-masing *medoid* menggunakan rumus jarak *euclidan.*
   c) Alokasi objek ke klaster dengan *medoid* terdekat, menentukan setiap Kecamatan masuk ke dalam klaster yang mana berdasarkan jarak paling dekat dengan *medoid.*
   d) Memilih *medoid* baru secara acak misalnya menggunakan data Kecamatan lain untuk menyatakan nilai setiap klaster.
   e) Menentukan total simpangan (S) dengan cara menghitung nilai total jarak baru dikurangi total jarak sebelumnya, yaitu S= TC_n. – TC_0. Jika $S > 0$, maka iterasi berhenti dan anggota klaster yang digunakan adalah anggota saat menghitung TC_0. Apabila S < 0, maka harus menukar objek dengan n data klaster sehingga terbentuk sekumpulan k objek dan dilakukan perulangan untuk langkah (c) hingga (d).
7. Analisis klaster menggunakan metode Fuzzy C-Means:
   a) Menentukan nilai w, MaxIter, dan ξ, menentukan nilai klaster yang sudah didapatkan melalui metode *elbow*, pangkat pembobot (w), maksimum iterasi (MaxIter) dan error terkecil (ξ) untuk menentukan kapan proses iterasi dapat berhenti.
   b) Membangkitkan matriks partisi awal $U_{nxc} = [\mu_{ik}]$, μ_{ik} yaitu bilangan *random* yang menyatakan suatu derajat keanggotaan.
   c) Menghitung pusat *cluster* ke-*k* (V_{kj}) dengan *k* = 1,2,…..,c; dan j = 1,2,….,m.
   d) Menghitung fungsi objektif pada iterasi ke-t, P_t yang menggambarkan jumlah jarak data ke pusat klaster.
   e) Menghitung perubahan matriks partisi dan dilakukan perbaruan sesuai hasil perhitungan.

f) Cek apakah kondisi berhenti tercapai, melakukan cek kondisi untuk berhenti atau tidak dengan ketentuan:
Jika ($|P_t - P_{t-1}| < \xi$) atau (t > MaxIter) maka berhenti;
Jika tidak: t = t+1, ulangi langkah (c).

8. Validasi hasil klaster dengan *silhouette coefficient* dari masing masing metode untuk menemukan *cluster* paling optimal.
9. Interpretasi karakteristik hasil klaster dari metode terbaik dari masing-masing klaster dengan membandingkan nilai rata-rata hasil iterasi maksimal dari setiap variabel dan dengan data rata-rata keseluruhan Kota Surabaya.

4. HASIL DAN PEMBAHASAN

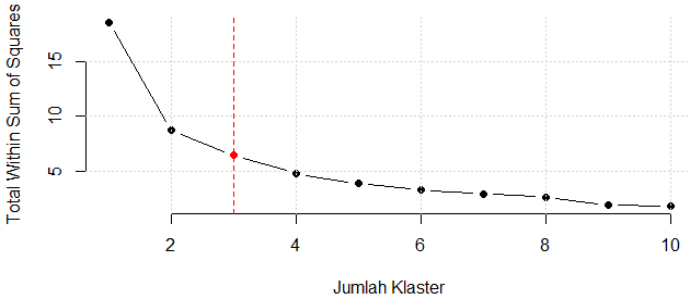
Dalam penelitian ini, data yang digunakan adalah data pendidikan pada 31 kecamatan di Kota Surabaya pada tahun 2022-2023. Proyeksi dari data tersebut dapat dilihat pada Tabel 2 berikut:

Tabel 2. Gambaran Umum Data Pendidikan Kota Surabaya Tahun 2022-2023

	Mean	Minimum	Maksimum
X_1	26.19	12	63
X_2	8446.67	3768	17778
X_3	445.19	200	916
X_4	12.19	5	25
X_5	3833.84	1380	7883
X_6	228.06	71	395
X_7	8.65	1	19
X_8	3917.97	512	12690
X_9	228.42	28	680

Uji kelayakan data pendidikan dilakukan menggunakan *Bartlett's Test of Sphericity* dan *Kaiser-Meyer-Olkin* (KMO) untuk memastikan relevansi atribut terhadap model. KMO dengan rentang nilai 0-1, menyatakan analisis faktor layak jika nilainya >0.5. Hasil perhitungan menunjukkan nilai KMO sebesar 0.68, sehingga analisis dapat dilanjutkan. Bartlett's test menghasilkan signifikansi 0.000 (<0.05), mengindikasikan bahwa atribut berkorelasi dan sesuai untuk proses selanjutnya.Data pendidikan yang layak dipakai tidak dapat langsung diolah karena adanya perbedaan besaran angka yang cukup jauh. Normalisasi data merupakan langkah penting dalam proses pengelompokan untuk memastikan bahwa setiap atribut memiliki skala yang sama dan tidak ada atribut yang mendominasi hasil akhir hanya karena skala yang berbeda.

Cara menentukan jumlah klaster (k) optimal pada penelitian ini menggunakan metode *elbow*. Peneliti telah menerapkan metode elbow pada data yang telah dinormalisasi, dan hasilnya berupa grafik seperti berikut:



Gambar 2. Grafik Metode *Elbow*

Berdasarkan Gambar 2 terjadi pergerakan grafik yang landai setelah pergerakan grafik yang curam atau berbentuk siku setelah titik 3. Dengan demikian, pada metode *elbow* didapatkan nilai k optimal adalah k=3.

Pengujian pertama dilakukan dengan metode *K-Means* menggunakan *k* optimal yang sudah didapatkan sebelumnya, yakni sejumlah tiga klaster yang kemudian dilakukan *clustering* berdasarkan posisi *centroid* terdekat. Proses iterasi berhenti ketika tidak terjadi perubahan pada *centroid* dan keseluruhan objek tidak dapat diklasifikasikan lagi, yaitu ketika iterasi ketujuh dan didapatkan hasil pada Table 3 berikut:

Tabel 3. Hasil *Clustering K-Means*

Klaster	Nama Kecamatan
Klaster 1	Karangpilang, Gayungan, Tenggilis Mejoyo, Gunung Anyar, Wiyung Tegalsari, Bulak, Pabean Cantian, Bubutan, Asemrowo, Benowo, dan Pakal.
Klaster 2	Wonocolo, Sukolilo, Mulyorejo, Gubeng, Wonokromo, Sukomanunggal, Sawahan, Genteng, Tambaksari, Kenjeran, Semampir, dan Krembangan.
Klaster 3	Jambangan, Rungkut, Dukuh Pakis, Lakarsantri, Sambikerep, Tandes, dan Simokerto.

Selanjutnya, untuk interpretasi karakteristik dari masing-masing klaster yang terbentuk maka dihitung rata-rata setiap atribut dalam klaster yang sama dan dibandingkan dengan rata-rata keseluruhan data Kota Surabaya, sehingga didapatkan hasil pada Tabel 4 berikut:

Tabel 4. Nilai Rata-rata Data Pendidikan Kota Surabaya dan Hasil Klaster K-Means

	Klaster 1	Klaster 2	Klaster 3	Surabaya
$\mathbf{X_1}$	0.159	0.563	0.279	0.342
$\mathbf{X_2}$	0.145	0.409	0.224	0.278
$\mathbf{X_3}$	0.149	0.523	0.254	0.334
$\mathbf{X_4}$	0.198	0.704	0.525	0.485
$\mathbf{X_5}$	0.129	0.506	0.386	0.360
$\mathbf{X_6}$	0.133	0.564	0.343	0.377
$\mathbf{X_7}$	0.125	0.434	0.210	0.307
$\mathbf{X_8}$	0.157	0.594	0.413	0.425
$\mathbf{X_9}$	0.113	0.385	0.177	0.280

Berdasarkan Tabel 4, Klaster 1 menunjukkan rata-rata jumlah guru, sekolah, dan murid yang lebih rendah dibandingkan rata-rata Kota Surabaya dan klaster lainnya, sehingga dapat dikategorikan sebagai wilayah dengan tingkat pendidikan rendah. Namun, pada tabel 8, rata-rata atribut Klaster 1 sebagian besar berada di atas rata-rata Kota Surabaya, kecuali pada X8 (jumlah murid SMA/SMK/MA) yang lebih rendah. Hal ini menempatkan Klaster 1 sebagai wilayah dengan tingkat pendidikan sedang namun memerlukan perhatian pada jenjang pendidikan SMA/SMK/MA. Sementara itu, Klaster 2 pada tabel 4 memiliki rata-rata atribut yang lebih tinggi dibandingkan klaster lain dan Kota Surabaya, mencerminkan tingkat pendidikan tinggi.

K-Medoids menggunakan *k* optimal yang sudah didapatkan sebelumnya, yakni sejumlah tiga klaster dan dilakukan *clustering* berdasarkan *medoid* terdekat yang dipilih ulang untuk setiap iterasi. Proses iterasi berhenti ketika setiap *medoid* baru yang dipilih tetap sebagai *medoid* setelah pertukaran titik, yaitu ketika iterasi keempat dan didapatkan hasil pada Tabel 5 berikut:

Tabel 5. Hasil *Clustering K-Medoids*

Klaster	Nama Kecamatan
Klaster 1	Karangpilang, Jambangan, Gayungan, Tenggilis Mejoyo, Gunung Anyar, Dukuh Pakis, Wiyung, Bulak, Simokerto, Pabean Cantian, Bubutan, Asemrowo, Benowo, dan Pakal.
Klaster 2	Wonocolo, Sukolilo, Mulyorejo, Gubeng, Wonokromo, Sambikerep, Sukomanunggal, Genteng, Tambaksari, Semampir, dan Krembangan.
Klaster 3	Rungkut, Lakarsantri, Tandes, Sawahan, Tegalsari, dan Kenjeran

Selanjutnya, untuk interpretasi karakteristik dari masing-masing klaster yang terbentuk maka dihitung rata-rata setiap atribut dalam klaster yang sama dan dibandingkan dengan rata-rata keseluruhan data Kota Suraba, sehingga didapatkan pada Tebel 6 berikut:

Tabel 6. Nilai Rata-rata Data Pendidikan Kota Surabaya dan Hasil Klaster *K-Medoids*

	Klaster 1	Klaster 2	Klaster 3	Surabaya
X_1	0.174	0.314	0.535	0.342
X_2	0.174	0.314	0.535	0.278
X_3	0.203	0.450	0.643	0.334
X_4	0.224	0.380	0.714	0.485
X_5	0.211	0.533	0.719	0.360
X_6	0.318	0.633	0.831	0.377
X_7	0.251	0.822	0.611	0.307
X_8	0.136	0.784	0.308	0.425
X_9	0.156	0.789	0.374	0.280

Berdasarkan Tabel 6 dapat dilihat bahwa nilai rata-rata pada klaster 1 selalu lebih rendah dibandingkan klaster lain dan data rata-rata Kota Surabaya. Dengan demikian, dapat disimpulkan Klaster 1 merupakan wilayah dengan jumlah guru, sekolah, dan murid berada di bawah rata-rata Kota Surabaya, yang menunjukkan bahwa klaster ini memiliki tingkat pendidikan yang rendah. Klaster 2 memiliki rata-rata yang lebih tinggi pada X_7-X_9 dibandingkan dengan keseluruhan Surabaya dan klaster lainnya. Hal ini menunjukkan bahwa klaster 2 memiliki jumlah guru, sekolah, dan murid yang tinggi untuk jenjang SMA/SMK/MA. Sementara itu, untuk . Klaster 3 memiliki rata-rata yang lebih tinggi pada X_1-X_6 dibandingkan dengan keseluruhan Surabaya dan klaster lainnya. Hal ini menunjukkan bahwa klaster 2 memiliki jumlah guru, sekolah, dan murid yang tinggi untuk jenjang SD/MI dan SMP/MTS.

Fuzzy C-Means menggunakan *k=3,* dilakukan *clustering* dengan memperbarui derajat keanggotaan fuzzy dan posisi *centroid* secara iteratif. Proses iterasi berhenti ketika perubahan nilai objektif lebih kecil dari nilai error (ξ) yang ditentukan di awal atau iterasi telah mencapai batas MaxIter yang ditentukan di awal juga. Pada perhitungan ini, proses iterasi berhenti saat iterasi ke-15 ketika nilai error (ξ) < 0.0001 dan didapatkan hasil pada table 7 berikut:

Tabel 7. Hasil *Clustering Fuzzy C-Means*

Klaster	Nama Kecamatan
Klaster 1	Wonocolo, Rungkut, Mulyorejo, Gubeng, Lakarsantri, Sambikerep, Tandes, Sukomanunggal, Sawahan, Semampir, dan Krembangan
Klaster 2	Karangpilang, Jambangan, Gayungan, Tenggilis Mejoyo, Gunung Anyar, Dukuh Pakis, Wiyung, Tegalsari, Genteng, Bulak, Simokerto, Pabean Cantian, Bubutan, Asemrowo, Benowo, dan Pakal
Klaster 3	Sukolilo, Wonokromo, Tambaksari, dan Kenjeran

Selanjutnya, untuk interpretasi karakteristik dari masing-masing klaster yang terbentuk maka dihitung rata-rata setiap atribut dalam klaster yang sama dan dibandingkan dengan rata-rata keseluruhan data Kota Suraba, sehingga didapatkan hasil pada table 8 berikut:

Tabel 8. Nilai Rata-rata Data Pendidikan Kota Surabaya dan Hasil Klaster *Fuzzy C-Means*

	Klaster 1	Klaster 2	Klaster 3	Surabaya
X_1	0.380	0.146	0.529	0.342
X_2	0.380	0.146	0.529	0.278
X_3	0.493	0.155	0.679	0.334
X_4	0.509	0.169	0.713	0.485
X_5	0.499	0.186	0.186	0.360
X_6	0.646	0.266	0.916	0.377
X_7	0.591	0.240	0.708	0.307
X_8	0.387	0.168	0.432	0.425
X_9	0.410	0.188	0.503	0.280

Pada Tabel 8, nilai rata-rata atribut Klaster 2 lebih rendah dibandingkan klaster lainnya dan rata-rata Kota Surabaya, sehingga dapat dikategorikan sebagai wilayah dengan tingkat pendidikan rendah. Untuk Klaster 3, tabel 4 menunjukkan rata-rata atribut berada di antara Klaster 1 dan 2, dengan beberapa nilai di atas atau di bawah rata-rata Surabaya, menggambarkan tingkat pendidikan sedang. Namun, tabel 8 menunjukkan rata-rata atribut Klaster 3 lebih tinggi dibandingkan klaster lain, mencerminkan tingkat pendidikan tinggi.

Sesudah melakukan pengklasteran, pada tahapan ini dilakukan validasi klaster menggunakan koefisien silhouette. Penggunaan koefisien silhouette berguna dalam melakukan validasi hasil klaster dari masing masing metode untuk menemukan klaster paling optimal sebagai langkah pemecahan masalah persebaran pendidikan di Kota Surabaya. Hasil nilai koefisien silhouette dengan bantuan R studio dihasilkan hasil seperti pada Tabel 9 berikut:

Tabel 9. Hasil Nilai Koefisien Silhouette

Metode	Nilai Rata-rata Koefisien *Silhouette*	Interpretasi
K-Means	0.592	*Good*
K-Medoids	0.257	*Weak*
Fuzzy C-Means	0.338	*Weak*

Berdasarkan Tabel 9, nilai rata-rata koefisien silhouette tiga metode yang digunakan, diperoleh nilai rata-rata 0,592 untuk metode *K-Means* diartikan ke dalam kriteria Good Classification dan nilai rata-rata 0,257 untuk metode *K-Medoids* diartikan ke dalam kriteria Weak Classification. Untuk *Fuzzy C-Means* memiliki nilai rata-rata 0,338 diartikan ke dalam kriteria Weak Classification. Dengan demikian, hasil klaster paling optimal dimiliki oleh metode *K-Means*.

5. KESIMPULAN

Dalam penelitian ini, metode terbaik yang diperoleh dari ketiga metode yang telah digunakan adalah *K-Means* dengan nilai rata-rata koefisien silhouette sebesar 0.592. Berdasarkan hasil analisis persebaran pendidikan di kota Surabaya menggunakan metode *K-Means* diperoleh sebanyak 12 kecamatan berada pada klaster 1, yakni Karangpilang, Gayungan, Tenggilis Mejoyo, Gunung Anyar, Wiyung, Tegalsari, Bulak, Pabean Cantian, Bubutan, Asemrowo, Benowo, dan Pakal. Kecamatan tersebut menunjukkan kondisi pendidikan yang rendah. Oleh karena itu, kecamatan-kecamatan tersebut memerlukan perhatian khusus pemerintah kota Surabaya untuk mengatasi kurangnya jumlah sekolah, jumlah murid, dan jumlah tenaga pendidik. Setelah menemukan hasil pengelompokan untuk persebaran pendidikan di Surabaya maka diharapkan nantinya dapat menjadi langkah yang tepat untuk mengatasi ketidakmerataan pendidikan di kota Surabaya.

DAFTAR PUSTAKA

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. International Journal of Computer Applications

Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. Matrix : Jurnal Manajemen Teknologi Dan Informatika

Haekal, M. (2022). Tantangan Distribusi Guru di Daerah Terpencil Indonesia: Antara Manajemen, Isu Personal, dan Faktor Geografi. TA'DIB : Jurnal Pemikiran Pendidikan

Hakim, L. (2016). Pemerataan Akses Pendidikan bagi Rakyat Sesuai dengan Amanat Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional. EduTech: Jurnal Ilmu Pendidikan Dan Ilmu Sosial, 2(1), 53–64.

Idris, F., Azmi, F., & Kusuma, D. P. (2019). Pengelompokan Data Guru Di Indonesia Menggunakan K-Means Clustering. E-Proceeding of Engineering, 6(2), 5643–5658.
Kurniawan, R., Mukarrobin, M., & Mahradianur. (2021). Klasterisasi Tingkat Pendidikan di DKI Jakarta pada Tingkat Kecamatan Menggunakan Algoritma K-Means. Technologia: Jurnal Ilmiah, 12(4), 234. https://doi.org/10.31602/tji.v12i4.5633
Madhulata, T. S. (2012). An Overview of Clustering Methods. IOSR Journal of Engineering
Mario, A., Herry, S., & Nasution, H. (2016). Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokkan Member Di Alvaro Fitness. Jurnal Sistem Dan Teknologi Informasi, 1(1), 1–6.
Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, & Pratama, E. A. (2018). Customer segmentation based on RFM model and clustering techniques with K-means algorithm. Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018, 1–6. https://doi.org/10.1109/IAC.2018.8780570
Meiriza, A., Ali, E., Rahmiati, & Agustin. (2023). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Program BPJS Ketenagakerjaan. Indonesian Journal of Computer Science, 12(2), 714–728. https://doi.org/10.33022/ijcs.v12i2.3184
Merliana, N. P. E., Ernawati, & Santoso, A. J. (2019). ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING. Matrix : Jurnal Manajemen Teknologi Dan Informatika, 102–109.
Nurfatimah, S. A., Hasna, S., & Rostika, D. (2022). Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs). Jurnal Basicedu, 6(4), 6145–6154. https://doi.org/10.31004/basicedu.v6i4.3183
Priambodo, Y. A., & Prasetyo, S. Y. J. (2018). Pemetaan Penyebaran Guru di Provinsi Banten dengan Menggunakan Metode Spatial Clustering K-Means (Studi kasus : Wilayah Provinsi Banten). Indonesian Journal of Computing and Modeling.
Putri, M. M., & Fithriasari, K. (2015). Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Metode Kohonen SOM dan K-Means. Jurnal Sains Dan Seni ITS, 4(1), 13–18. 10.12962/j23373520.v4i1.8815
Rachmatin, D. (2014). Aplikasi Metode-Metode Agglomerative dalam Analisis Klaster pada Data Tingkat Polusi Udara. Infinity Journal, 3(2), 133.
Reay, D. (2020). Miseducation: Inequality, Education, and the Working Classes. Journal of Teaching and Learning, 14(2), 64–66. https://doi.org/10.26522/ssj.v13i2.2227
Sari, P. I. (2019). Peran Pendidik dalam Implementasi Media Pembelajaran terhadap Peserta Didik Generasi 4.0. Prosiding Seminar Nasional Pendidikan FKIP, 2(1), 508–517.
Tasmalaila Hanifa, T., Al-Faraby, S., & Adiwijaya. (2017). Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging. E-Proceeding of Engineering, 4(2), 3210–3225.
Tchamyou, V. S. (2020). Education, lifelong learning, inequality and financial access: evidence from African countries. Contemporary Social Science, 15(1), 7–25.
Widarjono, A. 2010. Analisis Statistika Multivariat Terapan. Edisi pertama. Yogyakarta: UPP STIM YKPN.