

## EVALUASI K-MEANS DAN FUZZY C-MEANS CLUSTERING ANALISIS DATA HIV DI TINGKAT PROVINSI

Gabriel Arnold Budianto  
Teknik Informatika, Universitas Tarumanagara, Jakarta  
E-mail: \*[Gabriel.535210061@stu.untar.ac.id](mailto:Gabriel.535210061@stu.untar.ac.id)

### ABSTRAK

HIV (Human Immunodeficiency Virus) merupakan penyakit serius yang menyerang sistem kekebalan tubuh dengan menginfeksi darah dan sel-sel di dalam tubuh. Penyakit ini sangat berbahaya karena penyebarannya yang cepat, terutama melalui hubungan seksual berisiko, air liur, aktivitas menyusui oleh ibu yang terinfeksi, hingga penggunaan alat medis yang terkontaminasi seperti pada proses transfusi darah. Kota Surabaya di Provinsi Jawa Timur menjadi salah satu daerah dengan jumlah kasus HIV yang tinggi. Oleh karena itu, diperlukan analisis terhadap penyebaran kasus HIV di wilayah ini serta evaluasi layanan kesehatan yang diberikan berdasarkan standar nasional di berbagai puskesmas. Penelitian ini memanfaatkan data dari situs [katalog.data.co.id](http://katalog.data.co.id) dan menerapkan metode clustering K-Means serta Fuzzy-C-Means untuk mengevaluasi jumlah penderita HIV serta cakupan layanan kesehatan nasional yang mereka terima. Hasil dari pengelompokan ini dapat digunakan sebagai dasar pemetaan penyebaran HIV guna mencegah penyebaran lebih lanjut melalui tindakan penanganan yang tepat.

### Kata kunci

**Evaluasi, Clustering, HIV, K-Means, Fuzzy-C-Means**

### ABSTRACT

*HIV, or Human Immunodeficiency Virus, is a life-threatening disease that weakens the immune system by attacking the blood and body cells. Its rapid transmission, which can occur through unprotected sex, contact with infected saliva, breastfeeding, or contaminated medical equipment such as during blood transfusions, poses a serious health threat. Surabaya, located in East Java Province, is one of the regions with a high number of HIV cases. This study aims to analyze the distribution of HIV cases in Surabaya and assess treatment data aligned with Indonesia's national health standards using clustering techniques. Data was sourced from [katalog.data.co.id](http://katalog.data.co.id). The research applied K-Means and Fuzzy-C-Means clustering methods to determine the number of infected individuals and the extent to which they receive standardized healthcare services in each health center. The results of this clustering evaluation can help in mapping the spread of HIV and aid efforts to control and prevent its expansion.*

### Keywords

**Evaluation, Clustering, HIV, K-Means, Fuzzy-C-Means**

## 1. PENDAHULUAN

Human Immunodeficiency Virus (HIV) adalah virus yang menyerang sistem kekebalan tubuh, khususnya sel darah putih dan limfosit. Virus ini menyebabkan penurunan daya tahan tubuh, membuat penderitanya lebih rentan terhadap infeksi lain. Menariknya, beberapa individu dapat hidup dengan HIV tanpa menunjukkan gejala yang mencolok dan tidak selalu membutuhkan pengobatan langsung.

Antara tahun 2019 hingga 2023, HIV/AIDS menjadi perhatian publik karena banyak kasus terjadi pada anak-anak di bawah umur, khususnya di Surabaya, yang merupakan provinsi dengan jumlah kasus tertinggi kedua setelah DKI Jakarta dengan total 42.399 penderita. Penyebaran HIV di Indonesia banyak ditemukan di daerah perkotaan dan

Pemanfaatan teknologi seperti machine learning dapat menjadi solusi dalam menganalisis data penyebaran HIV. Sebuah studi sebelumnya oleh Rut Indra Lita Sinaga dan rekan-rekannya menggunakan metode K-Means untuk mengelompokkan jumlah kasus AIDS berdasarkan provinsi dari tahun 2016 hingga 2018.( Lesmana, A.A., et al., 2019) Metode ini menghasilkan dua kelompok: cluster rendah dan tinggi. Selain itu, metode Fuzzy-C-Means dapat digunakan untuk mengevaluasi seberapa besar pengaruh data terhadap setiap cluster melalui sistem keanggotaan.

## 2. METODE PENELITIAN

## 2.1 Sumber Data

### 2.1.2 Dataset

Dataset terdiri dari 304 entri yang mencakup fitur seperti kecamatan, kelurahan, jenis kelamin, bentuk penanganan, dan jumlah penderita HIV. Format data yang digunakan ditampilkan dalam Gambar 1.

[illegible]

**Gambar 1: Data Penderita HIV (Sumber:katalog.co.id)**

## 2. 2 Klasterisasi Data

Pada Penelitian ini menerapkan pendekatan *unsupervised learning* dengan metode klasterisasi untuk mengevaluasi dataset yang diambil dalam format CSV dari katalog.data.co.id. Dua algoritma utama digunakan dalam proses ini, yaitu K-Means dan Fuzzy C-Means, dengan evaluasi menggunakan indeks silhouette dan *Davies-Bouldin Index*. Klasterisasi bertujuan untuk mengelompokkan data mentah agar lebih mudah dianalisis dan dipahami.

## 2. 3 Metode K-Means

Algoritma K-Means bekerja dengan menentukan jumlah klaster (nilai K) yang diinginkan, lalu membagi data ke dalam klaster secara acak sebelum menghitung rata-rata untuk menentukan pusat klaster. K-Means merupakan metode sederhana yang digunakan untuk membagi entitas menjadi kelompok yang disebut klaster berdasarkan kesamaan karakteristik.( Nugroho, D.Y., Wibowo, A.Muhammadiyah, M., 2019). Algoritma ini menggunakan jarak Euclidean untuk mengukur kesamaan antar data, dengan rumus:

$$d(b_i, a_t) = \sqrt{\sum_{j=1}^l (b_{ij} - a_{tj})^2}$$

Keterangan :

- d adalah jarak antara data dengan pusat klaster
- Bi adalah nilai data
- at adalah pusat klaster

## 2. 4 Metode Fuzzy C-Means

metode Fuzzy C-Means (FCM) menggunakan konsep tingkat keanggotaan (membership degree) untuk setiap data terhadap setiap klaster. Berbeda dari K-Means yang hanya mengelompokkan data secara tegas, FCM memungkinkan suatu data memiliki keanggotaan di lebih dari satu klaster, sehingga lebih fleksibel dalam menangani data yang menyebar secara tidak teratur.

FCM sangat berguna ketika objek data memiliki karakteristik yang tumpang tindih antara beberapa klaster. Formula matematis dari metode ini melibatkan pusat klaster ( $V_{kj}$ ), derajat keanggotaan ( $\mu_{ik}$ ), dan nilai data ( $X_{ij}$ ) dengan rumus :

$$V_{kj} = \frac{\sum_{i=1}^n ((\mu_{ik})^w X_{ij})}{\sum_{i=1}^n (\mu_{ik})^w} ;$$

## 2. 5 Evaluasi Clustering Silhoutte Dan Davies-Bouldin Index

Evaluasi tambahan dilakukan menggunakan Clustering Silhoutte dan Davies-Bouldin Index untuk mengukur performa K-Means dan Fuzzy C-Means. Indeks ini mengevaluasi sejauh mana klaster saling terpisah dan seberapa kompak klaster tersebut. Semakin rendah nilai indeks ini, maka semakin baik hasil klasterisasi.

## Evaluasi Clustering: Silhouette

Nilai Silhouette	Interpretasi
0.71 - 1.00	Struktur yang dihasilkan kuat
0.51 - 0.70	Struktur yang dihasilkan baik
0.26 - 0.50	Struktur yang dihasilkan lemah
$\leq 0.25$	Tidak terstruktur

e

Experiment	Davies Bouldin Index Value					
	Simple K-Means			K-Means with centroid based min SSE		
	k=2	k=3	k=4	k=2	k=3	k=4
1	0.4296	0.0524	0.0406	0.1366	0.0493	0.0372
2	0.4296	0.0491	0.0358	0.2392	0.0540	0.0409
3	0.4296	0.0524	0.0322	0.2392	0.0505	0.0373
4	0.4296	0.0661	0.0367	0.2392	0.0493	0.0409
5	0.2845	0.0493	0.0391	0.2392	0.0493	0.0372
6	0.2845	0.0679	0.0515	0.2392	0.0493	0.0319
7	0.2845	0.0493	0.0320	0.2392	0.0493	0.0373
8	0.4296	0.0576	0.0407	0.2392	0.0491	0.0409
9	0.4296	0.0493	0.0365	0.1366	0.0493	0.0365
10	0.4296	0.0661	0.0407	0.2392	0.0493	0.0406
Average	0.38607	0.05595	0.03858	0.21868	0.04987	0.03807

**Gambar 2. Evaluasi Davies Bouldin Index Value**

### 3. HASIL DAN PEMBAHASAN

Pengelompokan data menggunakan algoritma K-Means menunjukkan bahwa seluruh data berhasil dikelompokkan berdasarkan jenis dan jumlah penderita HIV. Visualisasi hasil pengelompokan dan kurva *elbow* dari metode ini menggambarkan sebaran serta jumlah optimal klaster.

Pada metode Fuzzy C-Means (FCM), digunakan parameter *random\_state* yang mengontrol nilai acak dalam proses klasterisasi. Dengan menetapkan nilai awal acak, metode ini berhasil membentuk klaster sesuai penyebaran data, sebagaimana ditunjukkan dalam hasil visualisasi clustering.

Evaluasi performa kedua metode dilakukan menggunakan dua metrik, yaitu Silhouette Score dan Davies-Bouldin Index. Nilai *silhouette* menilai kualitas klaster berdasarkan keserupaan dalam klaster yang sama, sementara Davies-Bouldin mengukur seberapa baik klaster terbentuk berdasarkan jarak dan penyebaran antar klaster.

Hasil perbandingan menunjukkan bahwa secara umum, metode Fuzzy C-Means memberikan hasil yang lebih akurat dan representatif dibandingkan K-Means dalam mengelompokkan data penyebaran HIV di Surabaya. Evaluasi dari grafik dan indeks yang dihasilkan mengindikasikan bahwa FCM lebih sesuai untuk mendeskripsikan sebaran penderita HIV berdasarkan penanganan kesehatan yang sesuai standar nasional.

Pengelompokan data dengan metode K-Means terjadi dimana semua jumlah data per kategori sudah dikelompokkan berdasarkan jenis dan jumlah total data yang ada pada Gambar 4. Kemudian statistik data elbow yang muncul dapat dilihat pada Gambar 5.

```
[ ] Karakteristik rata-rata setiap cluster:
Cluster
kode_sarana_kesehatan lama_penanganan \
0.0 1.033304e+06 3.0
1.0 1.033316e+06 3.0
2.0 1.033326e+06 3.0

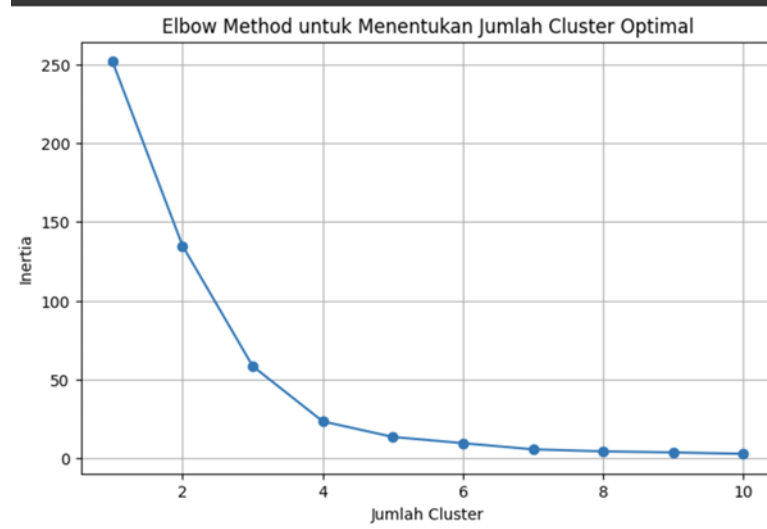
Cluster
jumlah_penduduk_terjangkit_laki2 \
0.0 1.0
1.0 0.0
2.0 0.0

Cluster
jumlah_penduduk_terjangkit_perempuan total_penduduk_terjangkit \
0.0 0.0 1.0
1.0 0.0 0.0
2.0 1.0 1.0

Cluster
jumlah_orang_terduga_menderita_hiv_yang_mendapatkan_pelayanan_kesehatan_sesuai_standar \
0.0 83.400000
1.0 56.631579
2.0 1.000000

Cluster
Cluster
0.0 0.0
1.0 1.0
2.0 2.0
```

**Gambar 3. Pengelompokan K-Means**



**Gambar 4. Elbow K-Means**

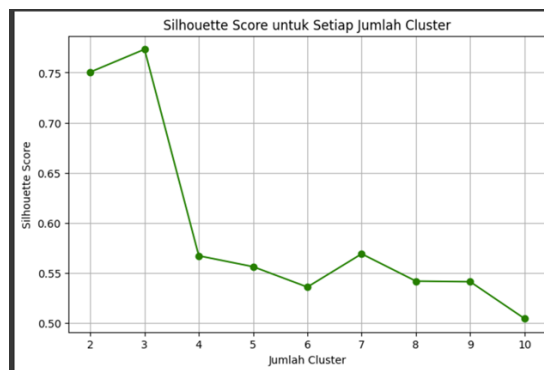
Selanjutnya peforma yang digunakan dalam metode ini yaitu Fuzzy C-Means hasil dapat dilihat pada Gambar 5.

```
Distribusi jumlah data per cluster:
Cluster
1.0 57
0.0 5
2.0 1
Name: count, dtype: int64
```

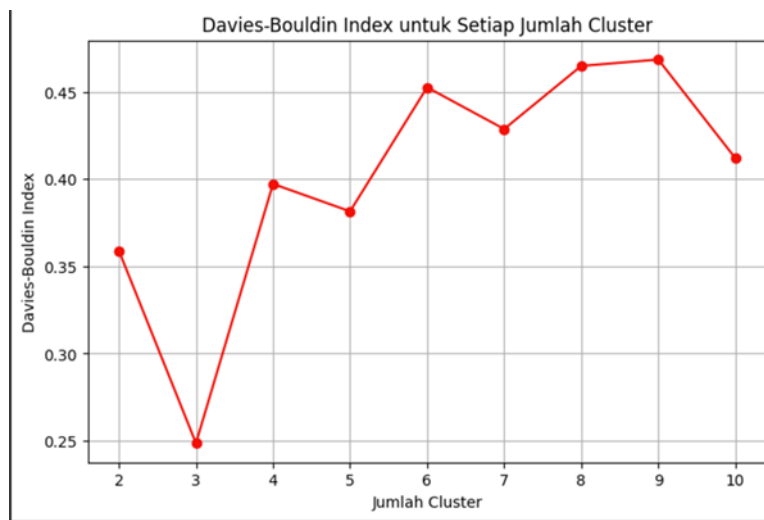
**Gambar 5. Hasil Clustering Fuzzy C-Means**

### 3.1 Evaluasi Clustering Silhoutte Dan Davies-Bouldin Index

Masing-masing metode ini melampirkan hasil elbow yang ada kemudian setelah di clustering data akan muncul Kembali dengan data yang sebenarnya dapat dilihat pada Gambar 6 metode silhouette dan Gambar 7 davies boildin



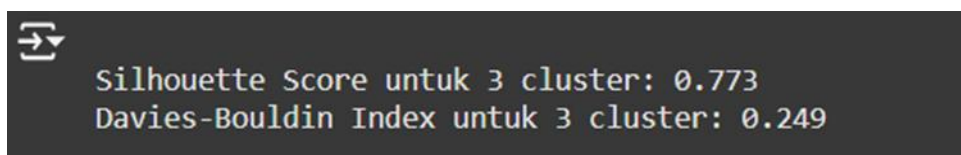
**Gambar 6. Silhouette Score**



**Gambar 7. Davies Boildin Index**

### 3.2 Perbandingan Peforma Algoritma

Metode clustering yang dilakukan dibandingkan akurasi dan evaluasi agar dapat terlihat metode mana yang memiliki performa terbaik. Perbandingan evaluasi clustering ini dapat dilihat pada Gambar 8.



**Gambar 8. Perbandingan Akurasi Clustering**

## 4. KESIMPULAN

Berdasarkan hasil evaluasi yang telah dilakukan, algoritma K-Means menghasilkan nilai *silhouette* sebesar 0,773 dengan pembentukan tiga klaster, sedangkan algoritma Fuzzy C-Means memperoleh nilai Davies-Bouldin Index sebesar 0,249 untuk jumlah klaster yang sama. Hasil tersebut mengindikasikan bahwa metode Fuzzy C-Means memiliki performa yang lebih baik dibandingkan K-Means dalam melakukan proses



klasterisasi data penyebaran HIV, khususnya yang berkaitan dengan standar penanganan kesehatan nasional di Provinsi Jawa Timur, Kota Surabaya. Meskipun demikian, diperlukan pengujian lanjutan dengan cakupan data yang lebih besar guna memperoleh hasil evaluasi yang lebih mendalam dan akurat.

## 5. DAFTAR PUSTAKA

- Alhamidi, R., 2022. Membaca Grafik Kasus HIV/AIDS di Jabar: Naik 3 Tahun Terakhir!. Detikjabar.[online] Available at: <https://www.detik.com/jabar/berita/d6262009/membaca-grafik-kasus-hiv-aids-dijabar-naik-3-tahun-terakhir> [Accessed 18 Jan. 2023].
- D. Nugroho, D.Y., Wibowo, A., Fakultas Kesehatan & Universitas Muhammadiyah, M., 2019. Analisis Cluster K-Means Kabupaten/Kota di Jawa Timur Berdasarkan Level Epidemik HIV. *Biometrika dan Kependudukan*, 8(2), pp.108–117. doi:10.20473/jbk.v8i2.2019.108.
- Dodu, A.Y.E., Nugraha, D.W. & Ayyub, M.A.A., 2019. Penerapan Data Mining untuk Mendeteksi Tingkatan Stadium Penyakit HIV/AIDS Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus pada RSUD Anutapura Palu). *Scientific: Computer Science and Informatics Journal*, 1(1), p.33. doi:10.22487/j26204118.2018.v1.i111900
- Herlinda, V. & Darwis, D., 2021. *Analisis Clustering untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means*. *Teknologi dan Sistem Informasi*, 2(2), pp.94–99. [online] Available at: <http://jim.teknokrat.ac.id/index.php/JTSL>.
- Lesmana, A.A., et al., 2019. Implementasi Algoritma K-Means untuk Clustering Penyakit HIV/AIDS di Indonesia. *eProceeding Engineering*, 6(2), pp.5564–5580.
- Noor, H., Dharmawati, A. & Qur'ana, T.W., 2021. Penerapan Algoritma K-Means Clustering Analysis pada Kasus Penderita HIV/AIDS (Studi Kasus Kabupaten Banjar). *Technologia: Jurnal Ilmiah*, 12(2), p.72. doi:10.31602/tji.v12i2.4573
- Rohmatullailah, D. & Fikriyah, D., 2021. *Faktor Risiko Kejadian HIV pada Kelompok Usia Produktif di Indonesia*. *Jurnal Biostatistik, Kependudukan, dan Informasi Kesehatan*, 2(1), p.45. doi:10.51181/bikfokes.v2i1.4652.
- Samudra, A.W.P., Susanto, R.A., Putra, A.R., Kurniadi, F.I. & Juarto, B., 2022. Klasifikasi HIV AIDS dengan Aplikasi Rapid Miner. *Sistem Komputer dan Kecerdasan Buatan*, (1), pp.15–19.
- Sinaga, R.I.L., Saputra, W. & Qurniawan, H., 2021. Pengelompokan Jumlah Kasus Penyakit AIDS Berdasarkan Provinsi Menggunakan Metode K-Means. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer Manajemen)*, 2(2), pp.99–107.