PERBANDINGAN ALGORITMA *K-MEANS*, K-*MEDOID*, DAN *DBSCAN* UNTUK *CLUSTERING* KUALITAS HIDUP INDONESIA DALAM PERSPEKTIF *KNOWLEDGE MANAGEMENT* DAN DATA DISCOVERY

Muhammad Ramadhan Putra Pratama, Muhammad Izzan Fieldi, Muhammad Syarief Albani , Muhammad Al Fachrozi , Fakih Rangga Aderiyana , Ken Ditha Tania, Allsela Meiriza
Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang - Prabumulih No.*KM*. 32, Indralaya Indah, Kec. Indralaya,
Kabupaten Ogan Ilir, Sumatera Selatan, *Indonesia*
muhammadramadhanputrapratam6@gmail.com

ABSTRAK

Kemajuan era digital mendunia memaksa manusia harus semakin peka dalam menggunakan teknologi dalam setiap aspek kehidupan. Khususnya pergerakan kualitas hidup di Indonesia, tantangan dalam menentukan metode yang paling baik untuk mengelompokkan wilayah berdasarkan indikator kualitas hidup. Permasalahannya adalah keberhasilan/keketepatan metode dalam melakukan clustering wilayah dengan karakteristik data yang berbeda-beda seperti pada kasus data yang kompleks, termasuk dataset yang mengandung outlier dan noise. Dampaknya pada efektivitas strategi pemerintah dan stakeholder yang mengimplementasikan post-strategi dengan memperhatikan hasilseling wilayah. Untuk itu, penelitian ini mengevaluasi metode clustering untuk memberikan representasi clustering yang lebih dekat dengan nilai aslinya dan efektivitas penerapannya pada masing-masing style dan design. Penelitian ini bertujuan untuk membanding kinerja dari tiga metode clustering, yaitu K-Means, Medoid Clustering, dan DBSCAN dalam mengelompokkan wilayah berdasarkan kualitas hidup di Indonesia. Prioritas penelitian ini adalah untuk menemukan beberapa metode clustering yang memberikan hasil yang lebih akurat, serta aplikatif untuk masing-masing hasil post-strateginya. Metode yang digunakan dalam penelitian ini adalah cluster analysis dengan tiga pendekatan utama, K-Means, Medoid Clustering, dan DBSCAN. Data dianalisis menggunakan teknik clustering yang berbeda dengan setiap metode clustering untuk mengevaluasi keefektifan, stabilitas, dan allevasi outlier dan noise. Metode yang digunakan dalam penelitian ini adalah cluster analysis dengan tiga pendekatan utama: K-Means, Medoid Clustering, dan DBSCAN. Data dikumpulkan dari berbagai sumber terkait indikator kualitas hidup, lalu dianalisis menggunakan teknik clustering untuk mengevaluasi efisiensi, stabilitas, dan ketahanan metode terhadap outlier serta noise. Hasil penelitian ini meningkatkan kontribusi pemerintah dalam memberikan strategi post bagi stakeholder berdasarkan hasil clustering. Integrasi kontribusi Knowledge Management dan Data Discovery dengan hasil clustering meningkatkan keefektifan analisis wilayah dalam luaran analisis strategi. Dalam studi selanjutnya, diharapkan memberikan data yang lebih akurat dengan menggunakan metode hybrid clustering atau machine learning.

Kata Kunci : *Data Mining; Knowledge Management, K-Means; K-Medoid; DBSCAN; Clustering.*

1. PENDAHULUAN

Indeks Pembangunan Manusia adalah indikator yang sesuai untuk menilai tingkat kemajuan atau keberhasilan suatu negara atau wilayah dalam pembangunan manusia. Indeks Pembangunan Manusia di Indonesia digunakan untuk mengukur tingkat kesejahteraan masyarakat melalui tiga indikator: angka harapan hidup untuk mencerminkan kualitas kesehatan, pendidikan diukur dengan rata-rata lama sekolah serta harapan lama sekolah, dan pendapatan per kapita sebagai indikator standar kehidupan [1].

Oleh karena itu, IPM dianggap lebih holistik karena hanya mengandalkan indikator ekonomi semata. Indonesia menduduki posisi keempat sebagai negara dengan jumlah penduduk terbesar di dunia [2] dan telah mengalami perubahan yang signifikan dalam aspek pembangunan manusia [3].

Sejak reformasi di akhir abad ke-20, pemerintah Indonesia telah berupaya untuk memperbaiki kualitas hidup warganya, sebagian besar melalui kebijakan pembangunan. Hal yang tidak kalah pentingnya adalah pengalokasian anggaran yang lebih besar ke sektor pendidikan dengan tujuan untuk memperluas akses ke pendidikan dan meningkatkan kualitas guru [4] serta reformasi di sektor kesehatan, dengan tujuan menjadi layanan yang lebih mudah diakses dan kualitas kesehatan Masyarakat [5].

Pencapaian ini dapat dilihat dari IPM Indonesia yang terus tumbuh positif setiap tahunnya [6].

Konsep *Knowledge Management* dan *Data Discovery* memiliki peran penting dalam pengelolaan informasi di era digital sekarang ini yang masih berkaitan dengan kualitas hidup. *Knowledge Management* menjadi fokus bagaimana data hasil pengumpulan informasi dapat diolah menjadi sebuah wawasan yang dapat dimanfaatkan oleh pemegang kebijakan untuk membuat keputusan yang lebih tepat. Sedangkan *Data Discovery* merupakan proses eksplorasi dalam analisis data sehingga memungkinkan penemuan pola terpendam dan hubungan antarkelompok dalam *dataset* kualitas hidup [7].

Dengan menggunakan konsep ini, kajian ini tidak hanya melakukan analisis *Clustering* tetapi juga mengintegrasikan *Knowledge Management* guna meningkatkan keberhasilan hasil riset dalam kebijakan publik..

2. TINJAUAN PUSTAKA

K-Means, *DBSCAN*, dan K-*Medoid* adalah metode *Clustering* tanpa pengawasan yang dapat digunakan untuk membagi data dalam beberapa *cluster* berdasarkan pola. *K-Means* bekerja dengan membagi *dataset* ke dalam *cluster* agar varians dalam *cluster* itu sendiri rendah, umumnya diberi kombinasi dengan algoritma Firefly untuk optimasi pengelompokan [8].

Sementara *DBSCAN* adalah singkatan dari *Density-Based Spatial Clustering of Applications with Noise*, mengelompokkan data yang *proximity spatial* dan pengabaian pada data yang jauh (*outlier*) [9].

K-*Medoid Clustering* bekerja dengan cara yang sama dengan *K-Means* tetapi lebih stabil dengan data *outlier* dan jauh lebih terakurat membentuk klise yang representatif; karena metode ini menggunakan titik data sebenarnya untuk klise *center* [10].

Dalam penelitian sebelumnya, diketahui bahwa K-*Medoid* membuat *cluster* yang lebih stabil dibandingkan dengan *K-Means*, terutama ketika melibatkan *dataset* dengan pola distribusi yang rumit [11].

Sejumlah penelitian terdahulu telah menggunakan *Clustering* untuk menganalisis data kualitas hidup dan lainnya. Contohnya penelitian oleh Aditya dkk. yang membahas Metode *K-Means* dan *DBSCAN* dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta [12].

Namun, penelitian-penelitian tersebut berfokus pada satu atau dua algoritma tertentu. Sampai saat ini, belum ada penelitian lengkap yang membandingkan ketiga algoritma *Clustering K-Means*, *Medoid Clustering*, dan *DBSCAN* dalam satu proyek khusus, khususnya sehubungan dengan kualitas hidup yang menggunakan data resmi dari BPS dalam bahasa Indonesia.

Selanjutnya, dalam penelitian Nahjan dkk., *RapidMiner* digunakan sebagai alat yang digunakan untuk mengelompokkan data transaksi *Oj Cell* dengan metode *K-Means Clustering*. Hasil perangkuman menunjukkan bahwa data telah dikelompokkan menjadi tiga *cluster* berdasarkan tingkat penjualan produk digital [13].

Penelitian yang lebih dalam menggunakan berbagai metode oleh karena itu, *RapidMiner* adalah alat yang sering digunakan di sebagian besar penelitian dalam data mining dan machine learning. *RapidMiner* sangat interaktif karena memposisikan penggunanya untuk memproses data, menerapkan berbagai algoritma, dan menganalisis set data untuk mendapat pola dan wawasan yang akurat [14].

Perangkat lunak ini juga fitur utama yang membantu dalam pemilihan algoritma serta memposisikan penggunanya untuk menerapkan model prediksi dalam tugas-tugas machine learning [15].

Tujuan dari penelitian ini adalah untuk membandingkan performa tiga jenis algoritma *Clustering*, yaitu *K-Means*, *Medoid Clustering*, dan *DBSCAN* pada pengelompokan data kualitas hidup di Indonesia tahun 2022-2024 yang optimal. Pembeda penelitian ini dengan penelitian sebelumnya adalah pendalaman analisis komparatif dari ketiga algoritma berdasarkan pada aspek akurasi, kecepatan, dan ketahanan terhadap *noise*. Hasil penelitian ini diharapkan dapat memberikan rekomendasi terhadap jenis algoritma *Clustering* yang tepat untuk analisis kualitas hidup serta kontribusi yang praktis bagi penanganan formulasi kebijakan publik yang lebih berbasis pada data.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistik Indonesia. Data tersebut berupa rata-rata Indeks Pembangunan Manusia pada periode 2022-2024 yang mencerminkan kualitas hidup masyarakat Indonesia berdasarkan indikator kesehatan, pendidikan, dan ekonomi di seluruh provinsi di Indonesia [16].

Setelah data diperoleh, data selanjutnya diimpor ke dalam perangkat lunak *RapidMiner* Studio untuk dianalisis.

Sebagaimana digunakan juga oleh [17] dalam membandingkan metode *K-Means*, K-*Medoid*, dan *DBSCAN* untuk menemukan hasil *cluster*isasi yang optimal. Diagram alur penelitian diawali dengan pengumpulan data IPM, kemudian preprocessing yang mencakup imputasi data kosong, normalisasi, serta penyesuaian atribut yang digunakan dalam analisis. Setelah data siap, maka dilakukan penerapan tiga algoritma *Clustering* yang berbeda yaitu *K-Means*, *Medoid Clustering*, serta *DBSCAN*.

Sebagai bagian dari analisis, pendekatan *Knowledge Management* dan *Data Discovery* diterapkan dengan tujuan untuk meningkatkan efektivitas pemrosesan data. *KM* berada pada sisi pengelolaan informasi dan memberikan hasil *Clustering* sebagai wawasan yang layak berguna dan bernilai untuk kegiatan perumusan kebijakan berbasis data. Di sisi lain, *Data Discovery* berusaha untuk membantu pengguna untuk menemukan pola-pola tersembunyi pada *dataset*. Dalam konteks kami, ini bisa berarti bahwa pola hubungan antara pendidikan dan kesejahteraan ekonomi di berbagai wilayah Indonesia tidak jelas karena kondisi *dataset* menjadi tidak cukup.

Adapun pada *K-Means*, jumlah *cluster* ditentukan menggunakan *elbow* dan jarak menggunakan *Euclidean Distance*. Dalam proses *Clustering* menggunakan *RapidMiner*, algoritma utama yang digunakan adalah operator *K-Means*. Parameter utama pada metode tersebut adalah jumlah *cluster* K yang diperbaharui dengan mengikuti metode

Elbow dan ukuran jarak yang dihitung dengan metode *Euclidean Distance*. Selain itu, jumlah iterasi maksimum yaitu 100 dikarenakan iterasi yang terlalu panjang dapat menyebabkan konvergen hasil yang optimal.

Algoritma *K-Means* merupakan metode *Clustering* yang bertujuan untuk membagi data ke dalam K kelompok berdasarkan kemiripan data. Parameter utama yang digunakan adalah:

a. Jumlah *cluster* (*K*)
b. Metode penghitungan jarak
c. Jumlah iterasi maksimum (umumnya 100 untuk menghindari iterasi yang terlalu panjang dan mengurangi risiko konvergensi lambat)

Rumus untuk menghitung jarak *Euclidean* antara dua titik data xx dan yy adalah:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Di mana:

a. $d(x,y)$= jarak *Euclidean* antara dua titik
b. x_i dan y_i= komponen ke-ii dari dua vektor data
c. n = jumlah dimensi fitur data [18].

Metode Elbow digunakan untuk menentukan nilai optimal dari K dengan cara mencari titik di mana penurunan inertia (jumlah kuadrat jarak tiap titik ke centroid) mulai melambat.

Sedangkan pada *Medoid Clustering*, *Manhattan Distance* digunakan untuk perhitungan jarak antar data. Dalam menggunakan operator K-*Medoid* di *RapidMiner*, terdapat beberapa parameter yang diatur agar hasil *cluster*isasi dapat dioptimalkan. Parameter utama yang perlu ditentukan adalah jumlah *cluster* atau K, yang menentukan seberapa banyak kelompok yang ingin dibentuk dalam proses *cluster*isasi. Selanjutnya, terdapat kriteria jarak yakni *Manhattan Distance* sebagai kriteria jarak yang digunakan, agar perhitungan jarak antara data yang digunakan untuk menentukan pusat *cluster* menjadi lebih optimal. Dalam pendekatan ini, K-*Medoid* akan menghasilkan *cluster* yang lebih stabil dan representatif terhadap data yang dianalisis. Algoritma K-*Medoid* mirip dengan *K-Means*, tetapi menggunakan *Medoid* (data aktual dalam *dataset*) sebagai pusat *cluster*, sehingga lebih tahan terhadap *outlier*. Parameter utama dalam algoritma ini adalah:

a. Jumlah *cluster* (*K*)
b. Metode penghitungan jarak (umumnya *Manhattan Distance*)

Adapun Rumus untuk menghitung jarak Manhattan antara dua titik data x dan y adalah:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Di mana:

a. $d(x,y)$ = jarak Manhattan antara dua titik
b. x_i dan y_i= komponen ke-ii dari dua vektor data
c. n = jumlah dimensi fitur data [19].

Pendekatan ini menghasilkan *cluster* yang lebih stabil dan representatif untuk data dengan *noise* atau *outlier* yang signifikan.

DBSCAN menggunakan dua parameter utama yaitu epsilon dan MinPoints yang diatur melalui eksperimen agar diperoleh *cluster* yang optimal. *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*). Dalam penggunaan algoritma *DBSCAN* di *RapidMiner*, ada satu parameter yaitu epsilon yang dirumuskan dalam ε serta *MinPoints* yang diharuskan untuk menentukan kepadatan data yang akan membentuk *cluster*. Nilai tepat dari parameter ini biasanya akan diperoleh dalam percobaan yang akan memberikan hasil *Clustering* yang lebih akurat . Selain itu, parameter yang tepat akan mempengaruhi banyak dan kualitas *cluster* yang dibuat pada akhir analisis data.

Algoritma *DBSCAN* didasarkan pada kepadatan data untuk membentuk *cluster* dan sangat efektif dalam mendeteksi data yang memiliki pola tidak beraturan. Parameter utama yang digunakan adalah:

a. ϵ*epsilon* (*epsilon*) — jarak maksimum yang dapat diterima untuk mengelompokkan titik-titik data dalam satu *cluster*.
b. MinPoints — jumlah minimum titik data yang harus ada dalam radius ϵ*epsilon* agar wilayah tersebut dianggap padat.

Adapun rumus penggunaan *DBSCAN* sebagai berikut. Jika D adalah himpunan data dan pp adalah titik data acak, maka:

$$N_\epsilon(p) = \{q \in D \mid d(p,q) \leq \epsilon\}$$

Keterangan:

a. $N_\epsilon(p)$ = himpunan semua titik yang berada dalam jarak ϵ*epsilon* dari titik pp
b. $d(p,q)$ = fungsi jarak (misalnya *Euclidean* Distance)

Jika jumlah titik dalam Nϵ(*p*)N_{*epsilon*}(p) lebih besar atau sama dengan MinPoints, maka titik pp menjadi pusat *cluster* [20].

Keseluruhan proses *Clustering* dilakukan menggunakan perangkat lunak *RapidMiner* Studio dan hasilnya divisualisasikan dalam tabel, scatter plot, serta dendrogram untuk menganalisis pola pergroupan. Evaluation *Clustering* dilakukan menggunakan dua metrik validasi internal, yaitu *Davies-Bouldin Index* serta *Silhouette Score*. Hal ini dilakukan untuk membandingkan performa ketiga algoritma *Clustering* tersebut dalam menghasilkan kelompok provinsi yang optimal dan menentukan metode yang efektif berdasarkan distribusi data dan tingkat heterogenitas antar wilayah.

4. HASIL DAN PEMBAHASAN

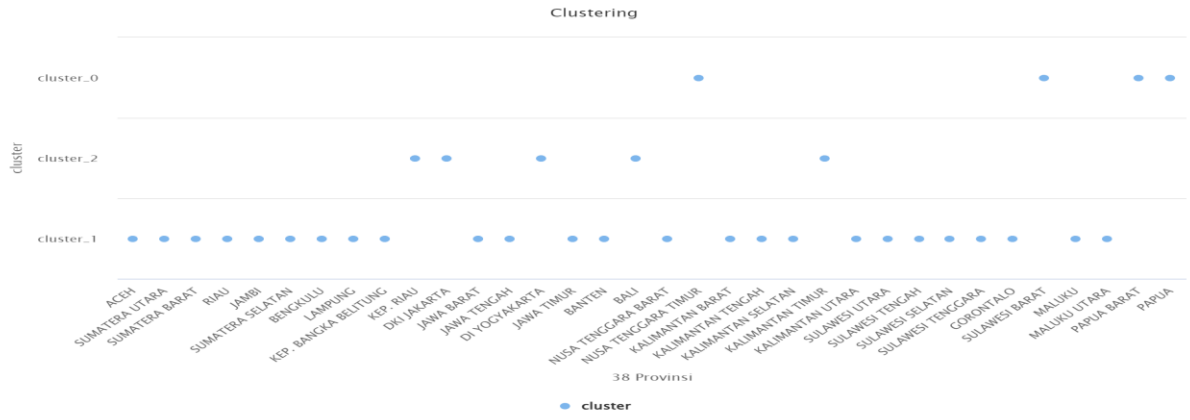
Penelitian ini mencoba untuk membandingkan performa tiga algoritma *Clustering*, yakni *K-Means*, *DBSCAN*, dan *Medoid Clustering*, pada *cluster* provinsi-provinsi di Indonesia berdasarkan indikator kualitas hidup. Dalam bayangan yang sudah dilakukan, masing-masing algoritma *Clustering* tersebut memunculkan pola pengelompokkan yang berbeda satu sama lain. Dengan demikian, analisis tersebut membuktikan bahwa setiap algoritma tersebut mempunyai keunikan karakteristik secara independen dalam menganalisa individualitas dari data provinsi tersebut.

4.1. Hasil *Clustering*

Tabel 1. Hasil *Clustering*

Provinsi	*Cluster K-Means*	*Cluster* K-*Medoid*	*Cluster DBSCAN*
Aceh	*cluster*_1	*cluster*_0	*cluster*_1
Sumatera Utara	*cluster*_1	*cluster*_0	*cluster*_1
Sumatera Barat	*cluster*_1	*cluster*_0	*cluster*_1
Riau	*cluster*_1	*cluster*_0	*cluster*_1
Jambi	*cluster*_1	*cluster*_0	*cluster*_1
Sumatera Selatan	*cluster*_1	*cluster*_0	*cluster*_1
Bengkulu	*cluster*_1	*cluster*_0	*cluster*_1
Lampung	*cluster*_1	*cluster*_0	*cluster*_1
Kep. Bangka Belitung	*cluster*_1	*cluster*_0	*cluster*_1
Kep. Riau	*cluster*_2	*Noise*	*cluster*_1
Dki Jakarta	*cluster*_2	*Noise*	*cluster*_1
Jawa Barat	*cluster*_1	*cluster*_0	*cluster*_1
Jawa Tengah	*cluster*_1	*cluster*_0	*cluster*_1
Di Yogyakarta	*cluster*_2	*Noise*	*cluster*_1
Jawa Timur	*cluster*_1	*cluster*_0	*cluster*_1
Banten	*cluster*_1	*cluster*_0	*cluster*_1
Bali	*cluster*_2	*Noise*	*cluster*_1
Nusa Tenggara Barat	*cluster*_1	*cluster*_0	*cluster*_1
Nusa Tenggara Timur	*cluster*_0	*Noise*	*cluster*_1
Kalimantan Barat	*cluster*_1	*Noise*	*cluster*_1
Kalimantan Tengah	*cluster*_1	*cluster*_0	*cluster*_1
Kalimantan Selatan	*cluster*_1	*cluster*_0	*cluster*_1
Kalimantan Timur	*cluster*_2	*Noise*	*cluster*_1
Kalimantan Utara	*cluster*_1	*cluster*_0	*cluster*_1
Sulawesi Utara	*cluster*_1	*cluster*_0	*cluster*_1
Sulawesi Tengah	*cluster*_1	*cluster*_0	*cluster*_1
Sulawesi Selatan	*cluster*_1	*cluster*_0	*cluster*_1
Sulawesi Tenggara	*cluster*_1	*cluster*_0	*cluster*_1
Gorontalo	*cluster*_1	*cluster*_0	*cluster*_1
Sulawesi Barat	*cluster*_0	*Noise*	*cluster*_1
Maluku	*cluster*_1	*cluster*_0	*cluster*_1
Maluku Utara	*cluster*_1	*cluster*_0	*cluster*_1
Papua Barat	*cluster*_0	*Noise*	*cluster*_1
Papua Barat Daya	*cluster*_0	*Noise*	*cluster*_0

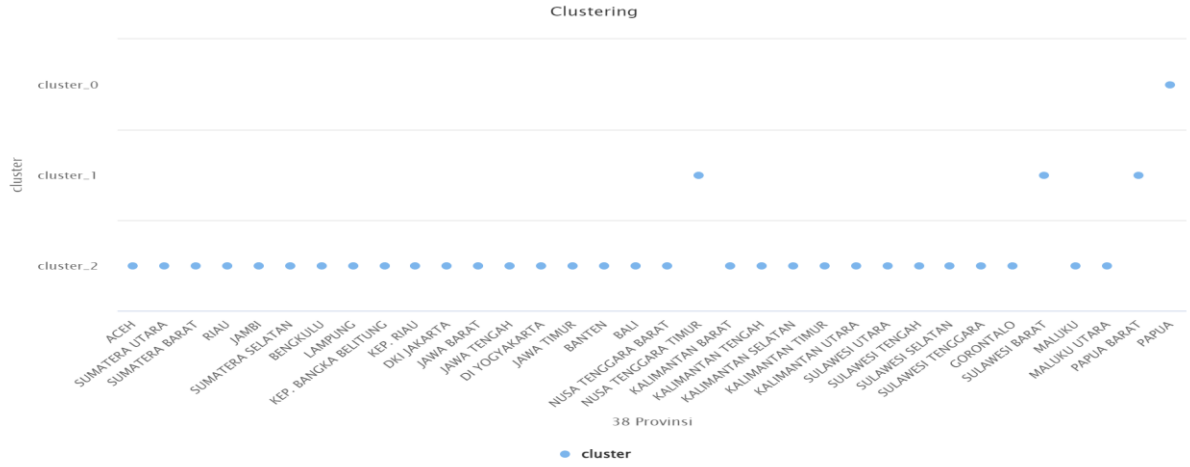
4.2. *K-Means Clustering*



Gambar 1. Visualisasi *K-Means*

Hasil dari pengaplikasian *Clustering* menggunakan metode *K-Means* menghasilkan ketiga *cluster* dengan distribusi yang lebih merata dari metode lainnya. *Cluster* 1 menjadi *cluster* terbanyak dengan 25 provinsi, yang diikuti oleh *Cluster* 0 dan *Cluster* 2 yang masing-masing terdiri dari 4 dan 5 provinsi. Karena datanya terdistribusi dalam ruang multidimensi berdasarkan kedekatan sentroid, *cluster* yang dihasilkan menjadi cukup rata dan ternilai optimal. Keunggulan dari metode ini adalah ketika distribusi data yang *cluster* optimal dilakukan pada data yang relatif homogen dan berbentuk konveks. Akan tetapi, kekurangan dari *K-Means* adalah ketergantungannya pada penentuan jumlah *cluster* yang dipilih, sehingga menentukan hasil akhir *cluster*isasi

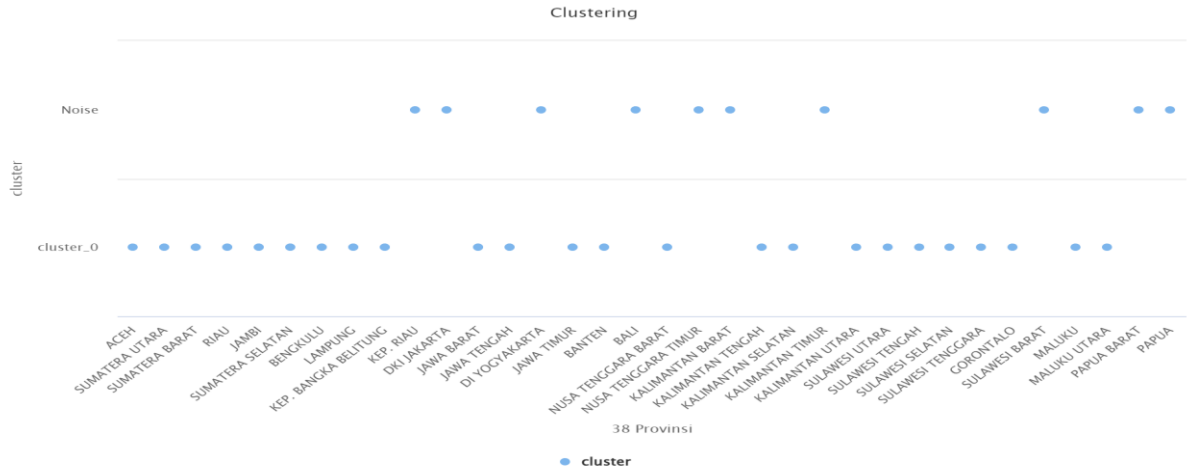
4.3. *K-Medoid Clustering*



Gambar 2. Visualisasi *K-Medoid*

Dari hasil tersebut. terlihat bahwa *cluster* yang terbentuk cenderung tidak seimbang, di mana *Cluster* 1 memiliki jumlah provinsi sebanyak 33 dan *Cluster* 0 hanya terdiri dari 1 provinsi. Hal ini menunjukkan bahwa algoritma *Medoid* cenderung kurang sensitif terhadap variasi antarkelompok. Hal ini terjadi karena algoritma ini mendeteksi representasi titik pusat berdasarkan total jarak minimum. Terlepas dari itu, *Medoid Clustering* memiliki keunggulan ketika melakukan pendekatan dengan data yang memiliki fitur bernilai ekstrim atau distribusi acak, namun hasil di atas menunjukkan bahwa algoritma ini tidak membentuk kluster-kluster yang optimal pada kasus ini. Masalah yang timbul, antara lain, adalah *lost of granularity* dalam pembentukan *cluster*; hal ini dapat menjadi kendala karena kita tidak bisa mengetahui seberapa besar heterogenitas kualitas hidup antar provinsi.

4.4. *DBSCAN (Density-Based Spatial Clustering of Application with Noise)*

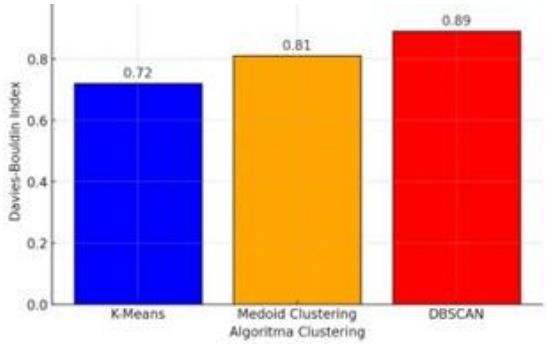


Gambar 3. Visualisasi *DBSCAAN*

Dibandingkan dengan *K-Means*, dalam *DBSCAN*, data dikelompokkan berdasarkan titik data area kepadatan di titik koordinat, didefinisikan oleh parameter minimum points dan epsilon sebagai *threshold* untuk kedekatan. Dari hasil *Clustering*, *DBSCAN* membagi data dalam dua *cluster* utama dan 10 provinsi yang merupakan *noise* yang disebut *outlier*. *Cluster* 0 merupakan *cluster* dengan data 24 provinsi dan sisanya tidak memiliki *cluster*. Dalam kasus ini, data mengindikasikan bahwa *DBSCAN* memiliki sensitivitas yang tinggi pada pola distribusi data yang heterogen. Jika pola distribusi data tidak homogen dan terdapat satu atau lebih *cluster* yang memiliki Provinsi dengan karakteristik yang berbeda jauh dari *cluster* lain, maka titik nilai tersebut cenderung di masukan menjadi satu kelas yaitu *noise*. Metode ini lebih cocok diterapkan dalam data yang memiliki pola distribusi yang heterogen dan tidak terstruktur atau banyak ditemui *outlier*. Metode ini lebih tidak optimal dalam kasus distribusi data terstruktur.

4.5. *Indeks Davies-Bouldin*

Evaluasi performa menggunakan *Silhouette Score*.

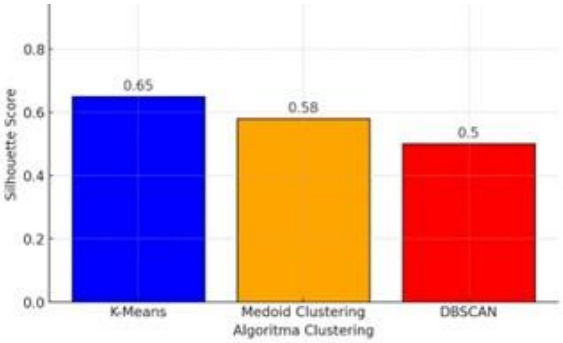


Gambar 4. Visualisasi Indeks Davies-Bouldin

Davies-Bouldin Index (DBI) digunakan untuk menilai seberapa baik suatu algoritma *Clustering* bekerja dengan melihat kedekatan antar *cluster* serta sebaran data di dalamnya. Semakin kecil nilai DBI, semakin baik hasil *Clustering* yang dihasilkan. Dari diagram batang, terlihat bahwa algoritma *K-Means* memiliki DBI terendah, yaitu 0.72. Ini menunjukkan bahwa *cluster* yang dihasilkan lebih terpisah dengan struktur yang lebih kompak dibandingkan *Medoid Clustering* (0.81) dan *DBSCAN* (0.89). Dengan kata lain, *K-Means* mampu memberikan hasil *Clustering* yang lebih optimal dalam konteks data yang digunakan.

4.6. *Silhouette Score*

Evaluasi performa menggunakan *Silhouette Score*.



Gambar 5. Evaluasi *Silhouette Score*

Silhouette Score digunakan untuk menilai seberapa baik hasil *Clustering* dengan melihat seberapa mirip suatu data dengan *cluster* tempatnya berada dibandingkan dengan *cluster* lain. Semakin tinggi nilainya, semakin baik kualitas *Clustering* yang dihasilkan. Dari diagram, terlihat bahwa *K-Means* memiliki nilai tertinggi, yaitu 0.65, diikuti oleh *Medoid Clustering* (0.58) dan *DBSCAN* (0.50). Ini menunjukkan bahwa *K-Means* mampu mengelompokkan data dengan lebih baik, sehingga setiap data berada dalam *cluster* yang paling sesuai dibandingkan dengan dua metode lainnya.

4.7. Integrasi *Knowlegde Management* dan *Data Discovery* dalam *Clustering*

Pendekatan *KM* dan *Data Discovery* diterapkan untuk mengeksplorasi hasil *Clustering* dan mengidentifikasi pola yang berhubungan dengan kesejahteraan masyarakat. Gambar di bawah berfokus pada bagaimana hasil *Clustering* menghubungkan dengan pendekatan *KM* dalam kualitas hidup data.

Gambar 6. Hubungan *Clustering* dan Penemuan Pola Data dalam Kualitas Hidup

Dengan pendekatan ini, hasil *Clustering* tidak hanya berfungsi untuk mengelompokkan wilayah, tetapi juga menjadi dasar bagi pengambilan keputusan berbasis data yang lebih efektif.

4.8. Perbandingan dan Implikasi

Perbandingan hasil *cluster*isasi berdasarkan tiga algoritma metode ini, menyoroti bahwa *K-Means* lebih efektif untuk menghasilkan distribusi *cluster* yang seimbang, sedangkan performa *DBSCAN* lebih baik untuk mendeteksi pola distribusi yang tidak terstruktur dan mengidentifikasi karakteristik provinsi yang berbeda sebagai *Noise*. Variasi algoritma *Medoid Clustering* yaitu hasil *cluster* yang tidak proporsional, paling banyak wilayah provinsi yang memiliki *cluster* dalam satu wilayah, ini justru berdampak pada sensitivitas variansi kualitas hidup antar-wilayah.

Selain pendekatan performa algoritma *Clustering*, selama proses penelitian ini, *Knowledge Management* dan *Data Discovery* memberikan perspektif lain kepada hasil *Clustering*. *KM* memberikan hasil *Clustering* menjadi bukan sekadar segmentasi data, tapi informasi yang dapat dikelola dan membantu pengambilan keputusan. Dengan bantuan *KM*, *cluster* hasil dapat diberikan faktor sosial-ekonomi dan faktor pendidikan, sehingga lebih memahami distribusi kualitas hidup di Indonesia.

Data Discovery memberikan peran pada pengetahuan bucket tersembunyi dalam data, di mana dapat membantu mendeskripsikan tren yang tidak ditemukan di aplikasi *Clustering* biasa. Contoh pengetahuan bucket adalah apabila memiliki kawasan tingkat pendidikan tinggi dan tingkat kesejahteraan ekonomi rendah, untuk membuat kebijakan terarah. Integrasi antara *KM* dan *Data Discovery* memberikan analisis *Clustering* yang tepat, dan membantu kebijakan berbasis bukti, sehingga dalam semua *cluster* dapat memberikan dampak ke publik.

Pada akhirnya, hasil ini memiliki dua implikasi besar terhadap dunia kebijakan. Pertama, dengan asumsi tujuan analisis adalah menemukan kelompok provinsi dengan karakteristik kualitas hidup yang seragam dalam distribusi yang relatif proporsional, kami lebih merekomendasikan algoritma *K-Means*. Disisi lain, dengan asumsi yang sama tetapi dengan tujuan analisis yang berlawanan, *DBSCAN* akan lebih direkomendasikan. Dan pada akhirnya, penggunaan *Medoid Clustering* kurang direkomendasikan, karena keseragaman hasilnya yang lemah dan kemampuan menggambarkan heterogenitas yang buruk antar wilayah.

Hasil dari penelitian ini menyarankan bahwa penelitian selanjutnya dapat mengoptimalkan parameter dan metodologi analisis prediksi, terutama metode hybrid, untuk dapat meningkatkan akurasi hasil dari kegiatan pengelompokkan. Beberapa parameter yang dapat disesuaikan termasuk nilai k dalam *K-Means* dan nilai epsilon serta MinPts dalam *DBSCAN*, yang dapat dicari untuk mendapatkan interpretasi data yang lebih baik berdasarkan berbagai kualitas hidup.

5. KESIMPULAN DAN SARAN

Dari hasil penelitian ini, terdapat kelebihan dan kekurangan dari metode tersebut dalam mengelompokkan wilayah berdasarkan kualitas hidup di Indonesia. *K-Means* sangat efisien dalam mengembangkan *cluster* dengan anggota yang maksimal, oleh karena itu akan lebih cocok untuk *dataset* kualitas hidup Indonesia. *Medoid Clustering* efisien dalam hal stabilitas *cluster* serta ketahanan terhadap *outlier* yang membuat teknik ini lebih unggul. *DBSCAN* dapat menemukan *noise* pada data dengan baik tetapi bekerja kurang efisien dalam menghasilkan lebih banyak *cluster* yang mirip untuk *dataset* ini.

Secara umum, dari hasil analisis distribusi *cluster* di atas, terlihat bahwa *K-Means* dan *Medoid Clustering* menghasilkan kelompok yang lebih terstruktur dibandingkan *DBSCAN*. Namun, dalam kegiatan penentuan metode terbaik, banyak faktor lain yang perlu dievaluasi, terutama karakteristik data dan tujuan analisis yang ingin dicapai. Karena itu, untuk memperoleh hasil optimal line, penerapan metode kualitas hidup secara realtime harus selalu menganalisis kualitas kebutuhan. Ini termasuk jumlah *cluster*, tingkat heterogenitas data, serta keberadaan *outlier* dan *noise*.

Diharapkan melalui penelitian ini dapat membantu pemerintah dan stakeholder untuk menetapkan strategi data dalam hal meningkatkan kualitas hidup. Dalam penelitian ini menggunakan metode *cluster* analysis seperti *K-Means Clustering* yang hanya mampu memberikan segmentasi kelistrikan daerah yang kurang akurat, maka penelitian lebih lanjut dapat menggunakan metode hybrid *Clustering* atau metode lanjutan dengan machine learning untuk mendapatkan segmentasi dan klasifikasi daerah terkait kualitas hidup yang lebih akurat lagi.

Selain itu, penerapan *Knowledge Management* (*KM*) dan *Data Discovery* dalam penelitian ini memungkinkan hasil *Clustering* tidak hanya menjadi sekadar segmentasi data, tetapi juga berfungsi sebagai wawasan strategis dalam analisis kualitas hidup. *KM* membantu mengorganisir dan mengelola hasil *Clustering* agar dapat dimanfaatkan lebih efektif dalam perumusan kebijakan berbasis data, sementara *Data Discovery* mengidentifikasi pola tersembunyi, seperti hubungan antara pendidikan dan kesejahteraan ekonomi. Dengan pendekatan ini, penelitian di masa depan dapat mengembangkan metode yang lebih prediktif dan adaptif.

DAFTAR PUSTAKA

[1] I. Ramadhan, B. Budiman, dan N. Alamsyah, "Optimization of Human Development Index in Indonesia Using Decision Tree C4.5, Support Vector Machine Algorithm, K-Nearest Neighbors, Naïve Bayes, and Extreme Gradient Boosting," *J. Inform.*, vol. 12, no. 1, hal. 15–26, Mar 2025, doi: 10.31294/inf.v12i1.21874.

[2] W. A. V. P. Aberth, "Indonesia Jadi Negara dengan Penduduk Terbanyak ke-4 di Dunia," data.goodstats.id. Diakses: 24 Maret 2025. [Daring]. Tersedia pada:

https://data.goodstats.id/statistic/indonesia-jadi-negara-dengan-penduduk-terbanyak-ke-4-di-dunia-DfGK7

[3] L. K. David, J. Wang, W. Brooks, dan V. Angel, "Digital transformation and socio-economic development in emerging economies: A multinational analysis," *Technol. Soc.*, vol. 81, hal. 102834, Jun 2025, doi: 10.1016/j.techsoc.2025.102834.

[4] R. Sparrow, T. Dartanto, dan R. Hartwig, "Indonesia Under the New Normal: Challenges and the Way Ahead," *Bull. Indones. Econ. Stud.*, vol. 56, no. 3, hal. 269–299, Sep 2020, doi: 10.1080/00074918.2020.1854079.

[5] N. Mboi *et al.*, "The state of health in Indonesia's provinces, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *Lancet Glob. Heal.*, vol. 10, no. 11, hal. e1632–e1645, Nov 2022, doi: 10.1016/S2214-109X(22)00371-0.

[6] Dr. Muchammad Romzi, "Indonesia's Human Development Index in 2023 reached 74.39, an increase of 0.62 points (0.84 percent) compared to previous year (73.77).," 2023. [Daring]. Tersedia pada: https://www.bps.go.id/en/pressrelease/2023/11/15/2033/indonesias-human-development-index-in-2023-reached-74-39--an-increase-of-0-62-points--0-84-percent--compared-to-previous-year--73-77--.html

[7] Novrizal Eka Saputra; Ken Ditha Tania; Rahmat Izwan Heroza, "Penerapan *Knowledge Management* system (*KM*s) menggunakan teknik knowledge *Data Discovery* (kdd) pada pt pln (persero) ws2jb rayon kayu agung," *J. Sist. Inf.*, vol. 8, no. 2, hal. 1038–1055, 2016.

[8] R. Chen, "Optimizing wireless sensor network topology with node load consideration," *Virtual Real. Intell. Hardw.*, vol. 7, no. 1, hal. 47–61, Feb 2025, doi: 10.1016/j.vrih.2024.08.003.

[9] D. K. Alfiki Astutik, A. Indrasetianingsih, dan F. Fitriani, "Penerapan Text Mining pada Analisis Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode Density Based Spatial *Clustering* of Applications With *Noise* (*DBSCAN*) dan *K-Means*," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 1, Jul 2022, doi: 10.36456/jstat.vol15.no1.a5983.

[10] Y. Saliba, A. Barbulescu, dan C. Ștefan Dumitriu, "A critical approach to *Clustering* precipitation series in the Dobrogea region, Romania," *E3S Web Conf.*, vol. 608, hal. 05027, Jan 2025, doi: 10.1051/e3sconf/202560805027.

[11] N. Dwitiyanti, Siti Ayu Kumala, dan Shinta Dwi Handayani, "Comparative Study of Earthquake *Clustering* in Indonesia Using K-*Medoid*, *K-Means*, *DBSCAN*, Fuzzy C-Means and K-AP Algorithms," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 6, hal. 768–778, Des 2024, doi: 10.29207/resti.v8i6.5514.

[12] A. Saputra dan R. Yusuf, "Perbandingan Algoritma *DBSCAN* dan *K-MEANS* dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta Menggunakan Metode RFM," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, hal. 1346–1361, Jul 2024, doi: 10.57152/malcom.v4i4.1516.

[13] M. Rafi Nahjan, Nono Heryana, dan Apriade Voutama, "IMPLEMENTASI *RAPIDMINER* DENGAN METODE *CLUSTERING K-MEANS* UNTUK ANALISA PENJUALAN PADA TOKO OJ CELL," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, hal. 101–104, Jan 2023, doi: 10.36040/jati.v7i1.6094.

[14] I. Domingues dan D. M. L. D. Rasteiro, "A Paradigm Shift in Teaching Machine Learning to Sustainable City Management Master Students," in *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, New York, NY, USA: ACM, Sep 2024, hal. 413–417. doi: 10.1145/3702163.3702448.

[15] M. Dehghani Saryazdi dan A. Mostafaeipour, "Identification and validation of key predictive factors for heart attack diagnosis using machine learning and fuzzy *Clustering*," *Eng. Appl. Artif. Intell.*, vol. 142, hal. 109968, Feb 2025, doi: 10.1016/j.engappai.2024.109968.

[16] Badan Pusat Statistik Indonesia, "[Metode Baru] Indeks Pembangunan Manusia menurut Provinsi, 2022-2024." [Daring]. Tersedia pada: https://www.bps.go.id/id/statistics-table/2/NDk0IzI=/-metode-baru-indeks-pembangunan-manusia-menurut-provinsi.html

[17] R. W. Sembiring Brahmana, F. A. Mohammed, dan K. Chairuang, "Customer Segmentation Based on RFM Model Using *K-Means*, K-*Medoid*, and *DBSCAN* Methods," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 11, no. 1, hal. 32, Apr 2020, doi: 10.24843/LKJITI.2020.v11.i01.p04.

[18] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, dan F. Foiadelli, "*K-Means* and Alternative *Clustering* Methods in Modern Power Systems," *IEEE Access*, vol. 11, hal. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.

[19] N. Fitrianti Fahrudin dan R. Rindiyani, "Comparison of K-*Medoid* and *K-Means* Algorithms in Segmenting Customers based on RFM Criteria," *E3S Web Conf.*, vol. 484, hal. 02008, Feb 2024, doi: 10.1051/e3sconf/202448402008.

[20] M. Raeisi dan A. B. Sesay, "A Distance Metric for Uneven *Cluster*s of Unsupervised *K-Means Clustering* Algorithm," *IEEE Access*, vol. 10, hal. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.