Jurnal Komputer dan Informatika
Vol 18 No 2, Oktober 2023: hlm 150 - 159

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

CLUSTERING DATA SISWA PUTUS SEKOLAH DENGAN ALGORITMA K-MEANS DAN DBSCAN

Mochamad Hammam Tegar Utomo
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: mochamad.535210054@stu.untar.ac.id,

ABSTRAK

Pembangunan sumber daya manusia dan kemajuan suatu negara bergantung pada pendidikan dasar. Meskipun partisipasi siswa di sekolah dasar telah meningkat pesat, fenomena putus sekolah di jenjang ini masih menjadi masalah besar di Indonesia. Penelitian ini menggunakan *algoritma clustering* K-Means dan DBScan untuk mengelompok data jumlah siswa putus sekolah setiap kota di Indonesia. Dataset yang digunakan berasal dari Kemendikbud yang diterbitkan pada tahun 2023, dataset ini memiliki variabel kota/kabupate, jumlah siswa putus sekolah SD, jumlah siswa putus sekolah SMP, jumlah siswa putus sekolah SMA, dan jumlah siswa putus sekolah SMK. Metode yang memiliki hasil terbaik yaitu algoritma K-Means dengan nilai K = 2 dengan nilai silhoutte 0.722. Secara umum hasil pengelompokkan menunjukkan daerah dengan jumlah siswa putus sekolah yang tinggi di angka 40 daerah, jumlah ini relatif rendah bila dibanding dengan kelompok daerah dengan jumlah siswa putus sekolah rendah yang bisa mencapai angka 474. Walaupun perbedaan yang cukup signifikan, hal ini bisa menandakan terdapat kesenjangan pendidikan antar daerah sehingga perbedaan nilai yang cukup jauh. Penelitian ini diharapkan bisa menjadi informasi penting bagi pemangku kepentingan.

Kata kunci—*K-Means, DBSCAN,* putus sekolah, pendidikan dasar

ABSTRACT

Human resource development and the progress of a country depend on primary education. Although the enrollment rate of students in primary school has increased rapidly, the dropout phenomenon at this level is still a major problem in Indonesia. This study uses K-Means clustering and DBScan algorithms to cluster data on the number of dropouts in each city in Indonesia. The dataset used comes from the Ministry of Education and Culture published in 2023. This dataset has city/district variables, the number of elementary school dropouts, the number of junior high school dropouts, the number of high school dropouts, and the number of vocational school dropouts. The method that has the best results is the K-Means algorithm with a value of K=2 with a silhouette value of 0.722. In general, the clustering results show that there are 40 regions with high dropout rates, relatively less than the 474 regions with low dropout rates. Although the difference is quite significant, it may indicate that there is an education gap between regions that results in a sizable difference in values. This research is expected to provide important information for stakeholders.

Keywords—*K-Means, DBSCAN, School Dropout, Primary Education*

1. PENDAHULUAN

Pembangunan sumber daya manusia dan kemajuan suatu negara bergantung pada pendidikan dasar. Meskipun partisipasi siswa di sekolah dasar telah meningkat pesat, fenomena putus sekolah di jenjang ini masih menjadi masalah besar di Indonesia. Anak-anak yang putus sekolah dasar berisiko menghadapi tantangan sosial dan ekonomi di masa depan, dan ini akan berdampak pada indeks pembangunan manusia yang rendah di seluruh negeri. Masalah ini seringkali dipengaruhi oleh faktor-faktor seperti kemiskinan, aksesibilitas pendidikan, dan kurangnya fasilitas sekolah, terutama di daerah terpencil. Oleh karena itu, sangat penting untuk memahami distribusi dan faktor-faktor yang memengaruhi tingkat putus sekolah di setiap provinsi untuk membuat kebijakan pendidikan yang efektif.

Untuk menganalisis pola dalam dataset pendidikan, ada beberapa teknik pengelompokan data, seperti K-Means dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). K-Means memungkinkan pengelompokan data ke dalam kelompok yang homogen berdasarkan kesamaan fitur, yang membantu menemukan wilayah dengan angka putus sekolah yang sebanding. Sebaliknya, DBSCAN memiliki kemampuan untuk mengidentifikasi pola pada data dalam bentuk cluster yang tidak beraturan dan menemukan data pencilan, atau outlier, yang dapat menunjukkan kasus atau anomali yang berbeda. Kedua metode ini dapat digunakan secara bersamaan untuk memberikan analisis yang lebih mendalam, karena masing-masing memiliki keunggulan dalam menangani karakteristik data yang berbeda. Kombinasi ini juga memungkinkan identifikasi pola tersembunyi yang sulit dideteksi dengan metode pengelompokan tunggal.

Tujuan dari studi ini adalah untuk menggunakan algoritma K-Means dan DBSCAN untuk menganalisis angka putus sekolah dasar di beberapa provinsi di Indonesia. Dataset yang digunakan mencakup informasi tentang provinsi, jumlah siswa yang meninggalkan sekolah dari kelas satu hingga kelas enam, dan kategori sekolah, baik negeri maupun swasta. Dengan menggunakan metode ini, penelitian diharapkan dapat mengelompokkan provinsi berdasarkan karakteristik angka putus sekolah dan menemukan provinsi dengan pola yang signifikan yang berbeda. Hasil pengelompokan ini diharapkan dapat memberikan wawasan yang berguna bagi pemerintah dalam menentukan prioritas alokasi sumber daya pendidikan. Selain itu, analisis ini juga dapat membantu mengidentifikasi faktor-faktor spesifik yang berkontribusi terhadap tingginya angka putus sekolah di daerah tertentu.

Studi sebelumnya menunjukkan bahwa metode pengelompokan data berguna dalam pendidikan. Sebagai contoh, penelitian yang menggunakan algoritma K-Means pengelompokan kabupaten di Jawa Tengah berdasarkan angka putus sekolah dasar; ini menghasilkan cluster dengan kategori tinggi, sedang, dan rendah [1]. Penelitian lain juga menggunakan algoritma ini untuk menganalisis tingkat partisipasi sekolah di berbagai provinsi di Indonesia, yang menunjukkan disparitas pendidikan antar wilayah [2].

Penelitian ini diharapkan dapat memberikan analisis yang lebih mendalam tentang angka putus sekolah dasar di Indonesia dengan menggabungkan metode K-Means dan DBSCAN. Penelitian ini diharapkan dapat membantu pembuat kebijakan menentukan daerah yang perlu diperhatikan dan strategi intervensi yang tepat untuk setiap kelompok. Tujuannya adalah untuk mengurangi angka putus sekolah dan mendorong pemerataan pendidikan di seluruh Indonesia. Selain itu, hasil penelitian ini diharapkan dapat menjadi acuan untuk pengembangan kebijakan berbasis data yang lebih akurat dan terukur. Dengan pendekatan yang tepat, langkah ini dapat mempercepat peningkatan kualitas pendidikan dan pembangunan sumber daya manusia di masa depan.

2. STUDI LITERATUR

2.1 Pendidikan Dasar

Pendidikan dasar adalah tahap awal dalam sistem pendidikan formal yang mencakup Sekolah Dasar (SD) dan Sekolah Menengah Pertama (SMP). Tahap ini bertujuan untuk memberikan pengetahuan dan keterampilan dasar kepada peserta didik, seperti membaca, menulis, berhitung, serta pengembangan karakter dan moral. Pendidikan dasar berperan penting dalam membentuk fondasi bagi pendidikan selanjutnya dan perkembangan individu secara keseluruhan [3].

Tujuan utama pendidikan dasar adalah membekali peserta didik dengan pengetahuan dan keterampilan dasar, menumbuhkan sikap yang diperlukan dalam masyarakat, serta mempersiapkan mereka untuk melanjutkan ke jenjang pendidikan menengah [4]. Pendidikan dasar juga berperan dalam membentuk kepribadian, akhlak mulia, dan keterampilan hidup mandiri pada peserta didik.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Pemerintah Indonesia mewajibkan program wajib belajar pendidikan dasar bagi setiap warga negara, yang menjadi tanggung jawab pemerintah. Hal ini bertujuan untuk mencerdaskan kehidupan bangsa dan mengembangkan potensi peserta didik agar menjadi manusia yang beriman, bertakwa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, serta menjadi warga negara yang demokratis dan bertanggung jawab.

Pendidikan dasar memainkan peran krusial dalam pembangunan sumber daya manusia dan kemajuan suatu negara, karena menjadi fondasi bagi perkembangan intelektual dan karakter individu. Oleh karena itu, kualitas dan aksesibilitas pendidikan dasar menjadi perhatian utama dalam upaya peningkatan kualitas pendidikan di Indonesia.

2.2 *Cluster*

Cluster adalah kumpulan objek atau data yang dikelompokkan bersama dalam satu grup berdasarkan karakteristik atau kesamaan tertentu, seperti atribut, pola, atau perilaku [5]. Dalam konteks ilmu data dan teknologi, klasterisasi memungkinkan pengelompokan data tanpa memerlukan label sebelumnya, sehingga berguna untuk menemukan pola atau struktur tersembunyi dalam dataset [6]. Konsep ini diterapkan secara luas di berbagai bidang, karena fleksibilitasnya untuk menangani beragam tipe data dan masalah. Beberapa contoh aplikasi meliputi:

llmu Data dan Statistik: Dalam analisis data, clustering adalah metode *unsupervised learning* yang mengelompokkan data ke dalam grup berdasarkan kemiripan atribut. Metode ini berguna dalam penemuan pola tersembunyi, segmentasi pasar, deteksi anomali, dan berbagai aplikasi lainnya. Ekonomi dan Bisnis: Dalam konteks ekonomi, cluster merujuk pada konsentrasi geografis dari perusahaan, pemasok, dan institusi terkait dalam industri tertentu. Cluster ekonomi dapat meningkatkan produktivitas dan inovasi melalui kolaborasi dan kompetisi sehat antar pelaku industri. Teknologi Informasi: Dalam komputasi, cluster komputer adalah kumpulan komputer yang terhubung dan bekerja bersama sebagai satu sistem untuk meningkatkan kinerja, ketersediaan, dan skalabilitas dalam pengolahan data dan layanan.

2.3 *K-Means*

K-Means adalah algoritma klasterisasi yang populer dalam bidang pembelajaran mesin dan analisis data [7]. Tujuan utama dari K-Means adalah mengelompokkan data ke dalam k *cluster* berdasarkan kesamaan fitur, di mana k adalah jumlah *cluster* yang ditentukan sebelumnya [8].
Cara Kerja K-Means:

1. Inisialisasi:
   Tentukan jumlah cluster k.
   Pilih k *centroid* awal secara acak atau menggunakan metode seperti K-Means++ untuk inisialisasi yang lebih efektif.
2. Langkah Penugasan (*Assignment Step*):
   Hitung jarak antara setiap data dan *centroid* menggunakan matrik jarak (misalnya, jarak Euclidean).
   Tetapkan setiap data ke cluster dengan centroid terdekat.
3. Langkah Pembaruan (*Update Step*):
   Setelah semua data ditugaskan, hitung ulang posisi centroid dengan mengambil rata-rata dari semua data dalam cluster tersebut.
4. Konvergensi:
   Ulangi langkah 2 dan 3 sampai posisi centroid tidak berubah signifikan atau setelah mencapai jumlah iterasi maksimum.

2.4 *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)*

DBSCAN adalah algoritma klasterisasi yang digunakan dalam pembelajaran mesin dan analisis data untuk mengelompokkan titik data berdasarkan kepadatan di sekitarnya [9]. Berbeda dengan metode seperti K-Means yang mengelompokkan data berdasarkan jarak ke centroid,

DBSCAN mengidentifikasi cluster sebagai area dengan kepadatan tinggi yang dipisahkan oleh area dengan kepadatan rendah.
Cara Kerja DBSCAN:

1. Parameter Utama:
   Epsilon (ε): Radius maksimum untuk mencari titik-titik tetangga dari sebuah titik data.
   Minimum Points (MinPts): Jumlah minimum titik dalam radius ε yang diperlukan agar sebuah titik dianggap sebagai titik inti (core point).
2. Jenis Titik:
   Core Point: Titik dengan setidaknya MinPts titik dalam radius ε.
   Border Point: Titik yang berada dalam radius ε dari sebuah core point tetapi memiliki kurang dari MinPts titik dalam radiusnya sendiri.
   Noise Point (Outlier): Titik yang bukan *core point* maupun border point.

2.5 *Elbow Method*

Elbow Method adalah teknik yang digunakan dalam analisis klaster untuk menentukan jumlah *cluster optimal* dalam algoritma seperti K-Means [10]. Metode ini membantu mengidentifikasi titik di mana penambahan jumlah cluster tidak lagi memberikan penurunan signifikan dalam *Within-Cluster Sum of Squares* [11].

Cara Kerja *Elbow Method*:

1. Menjalankan K-Means dengan Berbagai Nilai K:
   Jalankan algoritma K-Means berulang kali dengan jumlah cluster K yang berbeda, misalnya dari 1 hingga 10 atau lebih.
2. Menghitung WCSS untuk Setiap K:
   Untuk setiap nilai K, hitung total jarak kuadrat antara setiap titik data dan *centroid cluster*-nya.
3. Membuat Plot WCSS terhadap K:
   Buat plot dengan sumbu horizontal sebagai jumlah cluster K dan sumbu vertikal sebagai WCSS.
4. Menentukan Titik "Elbow":
   Identifikasi titik pada plot di mana penurunan WCSS mulai melambat secara signifikan. Titik ini menyerupai "siku" (elbow) dan dianggap sebagai jumlah cluster yang optimal.

2.6 *Within-Cluster Sum of Squares* (WCSS)

WCSS adalah ukuran yang digunakan untuk mengevaluasi seberapa baik data dikelompokkan dalam algoritma klasterisasi seperti K-Means [12]. WCSS mengukur total jarak kuadrat antara setiap titik data dengan centroid cluster tempat ia tergabung. Semakin kecil nilai WCSS, semakin baik data dikelompokkan karena ini menunjukkan bahwa titik-titik data berada lebih dekat dengan centroid cluster mereka.

$$WCSS = \sum_{k=1}^{k} \sum_{x \in c_k} \|x - \mu_k\|^2 \quad (1)$$

K: Jumlah cluster.
Ck: Cluster ke-k.
x: Titik data dalam cluster Ck.
μk: Centroid dari cluster Ck.
$\|x - \mu_k\|^2$: Jarak kuadrat Euclidean antara titik x dan centroid μk.

2.7 *Standard Scaler*

Standard Scaler adalah teknik penskalaan fitur yang disediakan oleh pustaka scikit-learn di Python [13]. Teknik ini menstandarkan fitur-fitur dataset Anda dengan menghapus rata-rata dan

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

menskalakan ke unit varians [14]. Proses ini memastikan bahwa semua fitur berkontribusi secara merata pada algoritma pengelompokan atau pembelajaran mesin, menghindari bias terhadap fitur dengan skala yang lebih besar.
For each feature in your dataset:

1. Compute the mean and standard deviation of the feature.
2. Transform each value in the feature using the formula:

$$z = \frac{x - \mu}{\sigma}$$

z: The standardized value.
x: Original value.
μ: Mean of the feature.
σ: Standard deviation of the feature

2.8 *Silhouette*

Silhouette adalah metode yang digunakan untuk mengevaluasi kualitas klasterisasi dalam algoritma seperti K-Means, DBSCAN, atau hierarki klaster [15]. Metode ini mengukur seberapa mirip titik data dengan kluster yang ditempatinya dibandingkan dengan kluster lainnya, memberikan indikasi seberapa baik data terklasifikasi ke dalam kluster masing-masing [16].
Untuk setiap titik data i, nilai Silhouette s(i) dihitung dengan langkah-langkah berikut:

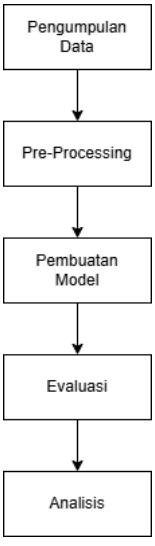
1. Hitung Rata-rata Jarak Dalam Kluster (a):
   a(i): Jarak rata-rata dari i ke semua titik lain dalam kluster yang sama.
2. Hitung Rata-rata Jarak ke Kluster Tetangga Terdekat (b):
   $b(i)$: Jarak rata-rata dari i ke semua titik dalam kluster tetangga terdekat (kluster berbeda dengan jarak rata-rata terkecil).

3. Hitung Nilai *Silhouette:*

$$s(i) = \frac{b(i) - a(i)}{maz(a(i), b(i))}$$

4. Rata-rata Semua Nilai Silhouette:
   Rata-rata s(i) untuk seluruh dataset memberikan ukuran kualitas klasterisasi secara keseluruhan.

3. METODE PENELITIAN



Gambar 1 Alur Penelitian

Alur penelitian ini dapat dilihat pada Gambar 1. Penelitian dimulai dengan pengumpulan *dataset*, *dataset* dikumpulkan dari situs Kemendikbud. Dataset dikumpulkan dari setiap kota di Indonesia, setiap provinsi memiliki data untuk setiap kelas. Data dapat diakses di https://katalog.data.go.id/dataset/jumlah-siswa-putus-sekolah. Pada dataset ini terdapat kolom daerah kota/kabupaten, jumlah putus sekolah jenjang SD, jumlah putus sekolah jenjang SMP, jumlah putus sekolah jenjang SMA, dan jumlah putus sekolah jenjang SMK. Dataset ini dikumpulkan pada tahun 2023 dengan total kota/kabupaten sebanyak 514 kota/kabupaten.

Tahap selanjutnya *Pre-processing* data, meliputi pembersihan data, pemilihan data, dan normalisasi data. Dalam tahap ini pembersihan data dilakukan pada data yang tidak diinginkan. Setelah itu data didalam dataset dinormalisasi menggunakan teknik normalisasi *standard scaler* untuk menyamakan nilai fitur agar hasil dari model menjadi lebih baik [17].

Model *clustering* yang digunakan yaitu K-Means dan DBScan. Setiap algoritma dilakukan percobaan dengan 10 skenario berbeda, skenario yang digunakan terdapat pada tabel 1. Dalam melakukan eksperimen akan menggunakan juga metode *elbow* dan *silhoutte analysis* untuk mencari nilai K terbaik.

Tabel 1 Skenario Uji Coba

Skenario	K-Means	DBScan
1	*Elbow Analysis*	epsilon = 0.5 dan min_samples = 2
2	*Silhoutte Analysis*	epsilon = 0.5 dan min_samples = 3
3	*K = 4*	epsilon = 0.5 dan min_samples = 4
4	*K = 5*	epsilon = 0.5 dan min_samples = 5
5	*K = 6*	epsilon = 0.5 dan min_samples = 6
6	*K = 7*	epsilon = 1.0 dan min_samples = 2
7	K = 8	epsilon = 1.0 dan min_samples = 3
8	K = 9	epsilon = 1.0 dan min_samples = 4
9	K = 10	epsilon = 1.0 dan min_samples = 5
10	K = 11	epsilon = 1.0 dan min_samples = 6

Setelah model dibuat, langkah selanjutnya yaitu mengevaluasi hasil dari model. Untuk melakukan evaluasi ini dilakukan perhitungan *Silhouette Score*. Metrik ini mengukur seberapa mirip suatu titik data dengan klaster yang menjadi anggotanya dibandingkan dengan klaster lainnya [18]
.

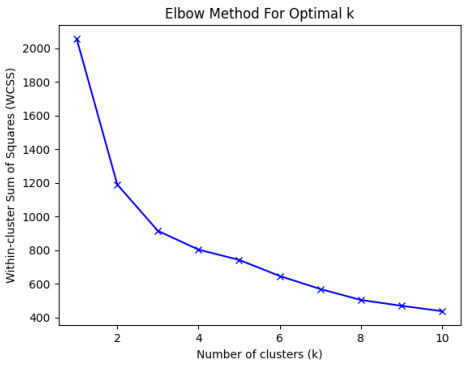
Silhouette memiliki rentang nilai 1-sampai dengan 1. Nilai *Silhouette* semakin tinggi menunjukkan hasil *clustering* semakin baik [19]. Jumlah *cluster* optimal atau jumlah *cluster* terbaik ditentukan berdasarkan nilai rata-rata *Silhouette* yang tertinggi [20]. Algoritma *clustering* yang memiliki nilai *Silhouette* semakin tinggi dapat dikatakan memeiliki kinerja yang baik. Hasil clusteringdengan nilai silhouettetertinggi kemudian dianalisis untuk menemukan informasi menarik tentang data putus sekolah di masing-masing *cluster*.

4.HASIL DAN PEMBAHASAN

Gambar 2 merupakan hasil eksperimen algoritma K-Means dengan menggunakan metode Elbow. Pada grafik elbow diatas WCSS turun tajam pada nilai K = 1 sampai dengan K = 3, setelah itu grafik mulai menurun landai. Pada analisis Elbow perlu dicari nilai K yang menurun tajam tetapi masih dekat dengan nilai K yang memiliki penurunan landai, dalam hal ini dapat diamati nilai K yang memenuhi kriteria tersebut terdapat pada K = 3. Selanjutnya akan dilakukan perbandingan nilai *Silhoutte Score* dengan skenario yang lain.

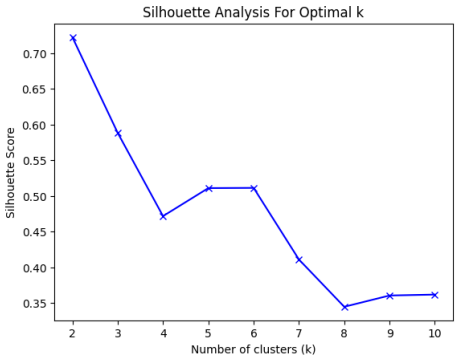
Gambar 2 merupakan hasil eksperimen algoritma K-Means dengan menggunakan metode Elbow. Pada grafik elbow diatas WCSS turun tajam pada nilai K = 1 sampai dengan K = 3, setelah itu grafik mulai menurun landai. Pada analisis Elbow perlu dicari nilai K yang menurun tajam tetapi masih dekat dengan nilai K yang memiliki penurunan landai, dalam hal ini dapat diamati nilai K yang memenuhi kriteria tersebut terdapat pada K = 3. Selanjutnya akan dilakukan perbandingan nilai Silhoutte Score dengan skenario yang lain

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)



Gambar 2 Visualisasi Elbow Analysis

Gambar 3 merupakan hasil eksperimen algoritma K-Means dengan menggunakan metode Silhoutte Analysis. Dari grafik tersebut, Silhoutte Score mengukur seberapa baik pemisahan cluster [21]. Silhoutte Score yang lebih tinggi mengindikasikan cluster yang terdefinisi dengan lebih baik [22]. Dalam grafik, Score Silhoutte tertinggi terdapat pada k = 2 selanjutnya terdapat penurunan yang cukup tajam dalam nilai Silhoutte dengan nilai dibawah 0.6 untuk k > 2. Berdasarkan grafik tersebut diambil nilai Silhoutte tertinggi yaitu pada K = 2.



Gambar 3 Visualisasi Silhoutte Analysis

Tabel 2 Nilai Silhoutte Untuk Setiap Skenario

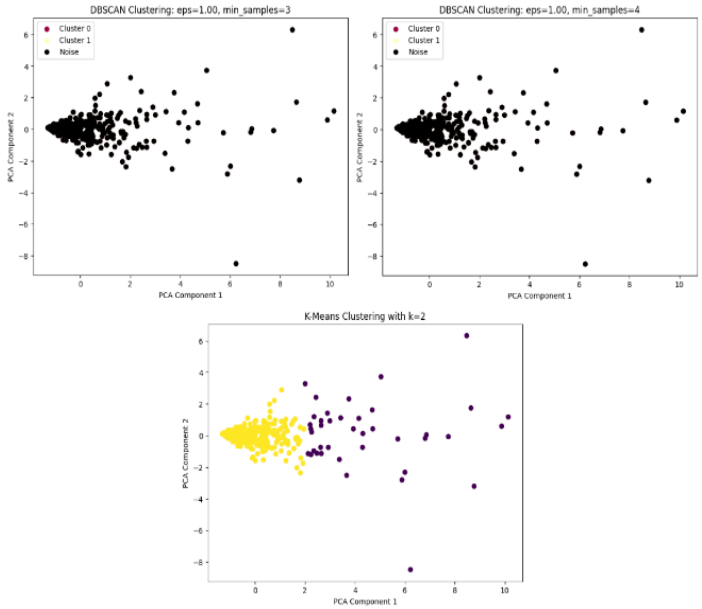
Skenario	K-Means	DBScan
1	*0.722*	0.247
2	*0.588*	0.272
3	*0.471*	0.276
4	*0.511*	N/A
5	*0.511*	N/A
6	*0.411*	0.529
7	0.344	0.538
8	0.360	0.538
9	0.362	0.524
10	0.367	N/A

Tabel 2 menampilkan hasil eksperimen dari skenario yang dirancang untuk mengevaluasi performa algoritma K-Means dan DBSCAN dalam klasterisasi data. Berdasarkan data pada tabel, algoritma K-Means menunjukkan performa terbaik dengan nilai *Silhouette* tertinggi sebesar 0.722. Nilai ini mencerminkan bahwa kluster yang dihasilkan oleh K-Means pada skenario ini memiliki kualitas yang sangat baik, dengan data yang lebih homogen dalam kluster dan jarak yang lebih jauh antar kluster.

Sebaliknya, pada algoritma DBSCAN, nilai *Silhouette* tertinggi ditemukan pada skenario 7 dan 8 dengan skor 0.538. Meskipun nilai ini lebih rendah dibandingkan K-Means, hasil tersebut tetap mengindikasikan bahwa skenario 7 dan 8 adalah konfigurasi terbaik untuk DBSCAN dalam

eksperimen ini. Kelebihan DBSCAN dalam mendeteksi noise dan outlier menjadikannya relevan dalam kasus di mana elemen-elemen tersebut memainkan peran penting, meskipun kualitas kluster secara keseluruhan mungkin tidak sebaik K-Means [23].

Dengan demikian, meskipun K-Means memberikan hasil yang lebih unggul dalam eksperimen ini, DBSCAN masih layak dipertimbangkan sebagai alternatif terutama ketika pengelolaan outlier atau identifikasi area dengan kepadatan rendah menjadi prioritas. Keputusan untuk menggunakan salah satu algoritma ini harus mempertimbangkan kebutuhan spesifik dari kasus atau dataset yang sedang dianalisis.



Gambar 4 Hasil Clustering K-Means dan DBScan

Pada Gambar 4 terdapat tiga grafik, yaitu hasil pengelompokkan dari DBScan Skenario 7, 8, dan K-Means dengan nilai K=2. Pada algoritma DBScan skenario 7 dihasilkan 2 kelompok berwarna kuning dan pink, dengan beberapa titik berwarna hitam untuk data yang dianggap noise. Grafik ini menangkap cluster berbasis kepadatan dan mengidentifikasi beberapa titik data sebagai *noise* yang tidak diperhitungkan oleh algoritma K-Means. Grafik kedua yaitu hasil dari DBScan skenario 8 memiliki kemiripan dengan grafik DBScan skenario 7 tetapi dengan jumlah titik hitam yang lebih banyak. Grafik ketiga merupakan hasil dari algoritma K-Means dengan K=2, pada grafik ini terdapat 2 kelompok dengan warna kuning dan ungu. Kelompok yang dibuat oleh algoritma K-Means memiliki perbedaan yang cukup jelas antar kelompok, sehingga bisa mendapatkan nilai *silhoutte* yang tinggi.

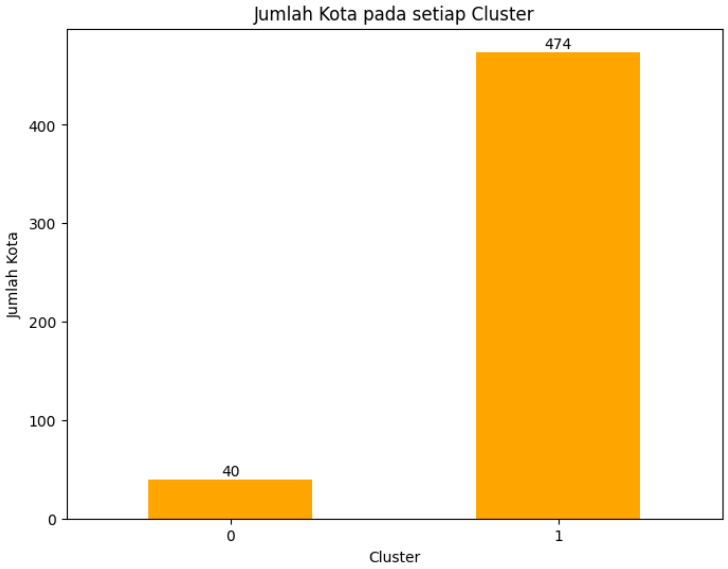
Tabel 3 Nilai Rata-Rata Putus Sekolah

Cluster	SD	SMP	SMA	SMK
0	289.100	115.775	87.175	132.925
1	61.151	19.149	13.898	14.951

Setelah mendapatkan hasil cluster terbaik yaitu algoritma K-Means dengan K=2, selanjutnya menganalisis hasil cluster dengan dataset. Pada Tabel 3 terdapat nilai rata-rata putus sekolah pada setiap cluster, cluster 0 memiliki nilai rata-rata putus sekolah tertinggi bila dibandingkan dengan cluster 1, perbedaaan nilai rata-rata hingga lebih dari 5 kali lipat untuk beberapa jenjang pendidikan.

Gambar 5 merupakan grafik jumlah kota pada setiap cluster, cluster 0 memiliki jumlah 40 dan cluster 1 memiliki jumlah 474. Dari gambar diatas menunjukkan jumlah putus sekolah tinggi hanya tersebar pada 40 kota/kabupaten saja, mayoritas daerah memiliki jumlah putus sekolah relatif rendah. Data tersebut dapat disimpulkan bahwa jumlah siswa putus sekolah disetiap daerah tidak terlalu tinggi, walaupun jumlah tersebut bisa dikurangin kembali kedepannya.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)



Gambar 5 Jumlah Kota Pada Setiap Cluster

5.KESIMPULAN

Kesimpulan dari penelitian ini yaitu menunjukkan potensi penggunaan algoritma K-Means dan DBSCAN untuk menganalisis angka putus sekolah pada pendidikan dasar di berbagai provinsi di Indonesia. Dengan menerapkan metode-metode ini, didapatkan metode terbaik untuk menganalisis dataset ini yaitu menggunakan algoritma K-Means dengan nilai K = 2. Dari algoritma tersebut didapatkan nilai *Silhoutte* tertinggi yaitu 0.722. Algorimat K-Means menghasilkan dua cluster, yaitu cluster 0 dan cluster 1. Cluster 0 merupakan daerah yang memiliki jumlah putus sekolah relatif lebih tinggi dibandingkan dengan cluster 1. Perbedaann rata-rata jumlah putus sekolah pada cluster 0 dengan cluster 1 hingga 5 kali lipat. Dari pengelompokkan ini didapatkan 40 daerah di Indonesia yang masih memiliki jumlah putus sekolah yang tinggi, dan 474 daerah memiliki jumlah putus sekolah yang lebih rendah. Hal ini memberikan kabar baik dalam pendidikan Indonesia. Temuan ini diharapkan dapat menjadi informasi penting bagi pemangku kepentingan dan pemerintah pusat dalam menekan jumlah siswa putus sekolah di Indonesia.

UCAPAN TERIMA KASIH

Jika ada, penulis dapat mengucapkan terima kasih kepada Alifia Dwinanti yang telah memberi dukungan terhadap penelitian ini. Jurnal Komputer dan Informatika berterima kasih kepada IJCCS sebagai landasan bentuk format makalah ini.

DAFTAR PUSTAKA

[1] Laila Khoirun Nisa, Tari Fitri Ningsih, Burhanuddin Izzul Salam, F. Fauzi, And Eny Winaryati, “Clustering Model K-Means Pada Kasus Angka Putus Sekolah Tingkatan Sekolah Dasar di Provinsi Jawa Tengah”, Logiclink, Vol. 1, No. 1, Hal. 13–20, Jun. 2024.

[2] Windarto, A.P., Herawan, T., K-Means Algorithm with Rapidminer in Clustering School Participation Rate in Indonesia. In: Ab. Nasir, A.F., Ibrahim, A.N., Ishak, I., Mat Yahya, N., Zakaria, M.A., P. P. Abdul Majeed, A. (Eds) Recent Trends in Mechatronics Towards Industry 4.0. Lecture Notes in Electrical Engineering, Vol 730, Hal. 779–794, 2022.

[3] Ade, Implementasi Kurikulum 2013 Dalam Pembelajaran SD/MI. Prenada Media, 2019.

[4] P. D. Purnasari and Y. D. Sadewo, 'Strategi Pembelajaran Pendidikan Dasar di Perbatasan Pada Era Digital', Jurnal Basicedu, Vol. 5, No. 5, Pp. 3089–3100, 2021.
[5] E. S. Dalmaijer, C. L. Nord, And D. E. Astle, Statistical Power for Cluster Analysis', BMC Bioinformatics, Vol. 23, No. 1, P. 205, 2022.
[6] Kais Ghedira Et Al., "Design and Implementation of a Scalable High-Performance Computing (HPC) Cluster for Omics Data Analysis: Achievements, Challenges and Recommendations in Lmics.," Gigascience, Vol. 13, Jan. 2024, Doi: Https://Doi.Org/10.1093/Gigascience/Giae060.
[7] [1] Priyanka Nandal, Optimizing Web Search Results for Image. K-Means Clustering Algorithm. GRIN Verlag, 2021.
[8] João Moreira, C. Ponce, And Tomáš Horváth, A General Introduction to Data Analytics. Chichester: Wiley Blackwell, 2019.
[9] Tshepo Chris Nokeri, Data Science Revealed: Feature Engineering, Data Visualization, Pipeline Development, And Hyperparameter Tuning. United States: Apress, 2021.
[10] B. Peter, DATA MINING BUSINESS ANALYTICS: Concepts, Techniques and Applications in Python. S.L.: Wiley-Blackwell, 2020.
[11] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," 2020, Doi: Https://Doi.Org/10.23977/Accaf.2020.010102.
[12] S. Priya and R. Manavalan, "Kmeans-NM-Salpepi: Genetic Interactions Detection Through K-Means Clustering with Nelder-Mead and Salp Optimization Techniques in Genome-Wide Association Studies," Artificial Intelligence Evolution, Pp. 67–80, Oct. 2021, Doi: Https://Doi.Org/10.37256/Aie.2220211099.
[13] F. Pedregosa Et Al., 'Scikit-Learn: Machine Learning in Python', Journal of Machine Learning Research, Vol. 12, Pp. 2825–2830, 2011.
[14] J. Brownlee, Data Preparation for Machine Learning. Machine Learning Mastery, 2020.
[15] K. Jajuga, BatógJ., And M. Walesiak, Classification and Data Analysis: Theory and Applications. Cham: Springer, 2020.
[16] S. Paembonan and H. Abduh, Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat', PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik, Vol. 6, P. 48, 09 2021.
[17] C. Fan, M. Chen, X. Wang, J. Wang, And B. Huang, 'A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data', Frontiers in Energy Research, Vol. 9, P. 652801, 2021.
[18] K. R. Shahapure and C. Nicholas, 'Cluster Quality Analysis Using Silhouette Score', In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, Pp. 747–748.
[19] M. Shutaywi and N. N. Kachouie, Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering', Entropy, Vol. 23, No. 6, P. 759, 2021.
[20] K. R. Shahapure and C. Nicholas, 'Cluster Quality Analysis Using Silhouette Score', In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, Pp. 747–748.
[21] A. Dudek, 'Silhouette Index as Clustering Evaluation Tool', In Classification and Data Analysis: Theory and Applications 28, 2020, Pp. 19–33.
[22] M. Shutaywi and N. N. Kachouie, Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering', Entropy, Vol. 23, No. 6, P. 759, 2021.
[23] D. Deng, 'DBSCAN Clustering Algorithm Based on Density', In 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), 2020, Pp. 949–953.
[24] J. Brownlee, Data Preparation for Machine Learning. Machine Learning Mastery, 2020.
[25] K. R. Shahapure and C. Nicholas, 'Cluster Quality Analysis Using Silhouette Score', In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, Pp. 747–748