

ANALISIS PADA TINGKAT PENYELESAIAN PENDIDIKAN DI 34 PROVINSI INDONESIA MENGGUNAKAN METODE ALGORITMA K-MEANS CLUSTERING

Noeralita Vefiliana Audya Wulandari*, Guruh Wijaya, Habibatul Azizah Al-Faruq

Program Studi Teknik Informatika, Universitas Muhammadiyah Jember

Email: noeralitavefiliana@gmail.com*, gruh.wijaya@unmuhjember.ac.id, habibatulazizah@unmuhjember.ac.id

ABSTRAK

Pendidikan merupakan salah satu komponen utama yang menentukan kemajuan suatu negara. Namun, ada perbedaan dalam tingkat pendidikan di berbagai provinsi menunjukkan adanya disparitas, atau perbedaan, yang perlu diteliti secara menyeluruh untuk mengidentifikasi penyebabnya dan menemukan solusi yang tepat. Penelitian ini menganalisis tingkat penyelesaian pendidikan di 34 provinsi di Indonesia menggunakan algoritma *K-Means Clustering*. Data yang digunakan berasal dari Badan Pusat Statistik Indonesia untuk periode 2021–2023. Proses pengelompokan dilakukan untuk jenjang pendidikan SD, SMP, dan SMA dengan bantuan perangkat lunak *RapidMiner*. Hasil analisis menunjukkan bahwa terdapat pola berbeda di setiap jenjang pendidikan, di mana *cluster* dengan nilai tertinggi untuk jenjang SD, SMP, dan SMA masing-masing memiliki nilai *Davies-Bouldin Index* (DBI) sebesar 0,393, 0,489, dan 0,562, yang mencerminkan kualitas klasterisasi cukup baik. Penelitian ini diharapkan dapat menjadi dasar bagi pengambilan kebijakan pendidikan untuk mengurangi disparitas tingkat pendidikan di Indonesia.

Kata Kunci: *Data Mining, K-Means Clustering, Tingkat Penyelesaian Pendidikan, Cluster*

ABSTRACT

Education is one of the key components that determine a country's progress. However, differences in educational levels across various provinces indicate disparities that need to be thoroughly studied to identify their causes and find appropriate solutions. This study analyzes the educational completion rates in 34 provinces of Indonesia using the *K-Means Clustering* algorithm. The data used were obtained from the Central Bureau of Statistics of Indonesia for the period 2021–2023. The clustering process was conducted for elementary, middle, and high school levels with the assistance of *RapidMiner* software. The analysis results revealed distinct patterns at each educational level, where the clusters with the highest scores for elementary, middle, and high school levels achieved *Davies-Bouldin Index* (DBI) values of 0.393, 0.489, and 0.562, respectively, indicating fairly good clustering quality. This study is expected to serve as a foundation for educational policy-making to reduce disparities in educational levels across Indonesia.

Keywords: *Data Mining, K-Means Clustering, Educational Completion Rate, Cluster*

1. PENDAHULUAN

Salah satu komponen utama yang menentukan kemajuan suatu negara adalah pendidikan. Pemerintah Indonesia telah mengambil beberapa kebijakan untuk meningkatkan akses dan kualitas pendidikan di berbagai jenjang, dari sekolah dasar hingga perguruan tinggi. Namun, perbedaan dalam tingkat pendidikan di berbagai provinsi menunjukkan adanya disparitas, atau perbedaan, yang perlu diteliti secara menyeluruh untuk mengidentifikasi penyebabnya dan menemukan solusi yang tepat.

Di era internet saat ini, teknik *data mining* untuk analisis pendidikan menjadi semakin penting. Algoritma pengelompokan *K-Means Clustering*, yang memungkinkan pengelompokan data berdasarkan atribut tertentu, adalah salah satu teknik yang dapat digunakan. Proses pengelompokan ini dapat dilakukan secara efisien dan efektif dengan menggunakan alat seperti *RapidMiner*, yang menyediakan *platform* untuk analisis data yang komprehensif.

Penelitian sebelumnya, yang dilakukan oleh (Indriyani & Irfiani, 2019) berjudul *Clustering Data Penjualan Pada Toko Perlengkapan Outdoor* menggunakan Metode *K-Means Clustering*, mengungkapkan bahwa algoritma *K-Means Clustering* memiliki kemampuan untuk meningkatkan kepuasan konsumen dengan mengelompokkan data persediaan produk penjualan. Toko Genta Corp mengalami kesulitan dalam menentukan jumlah produk yang tersedia dikarenakan pencatatan data penjualan yang masih dilakukan secara manual. Akibatnya, terdapat produk yang kurang laris dalam persediaan dan produk yang habis. Nilai *centroid* c1 adalah 64,7% dari produk yang paling laris,

centroid c2 adalah 24,3% dari produk yang cukup laris, dan *centroid* c3 adalah 24,3% dari produk yang cukup laris. Semua ini dihasilkan dengan menggunakan algoritma *K-Means Clustering*. *Centroid* dengan nilai tertinggi, yaitu c1 sebesar 11,1%, mengindikasikan produk yang kurang laris, sedangkan *centroid* c2 sebesar 24,3% mengindikasikan produk yang cukup laris. Data ini dapat digunakan oleh manajemen untuk mengelola persediaan produk secara lebih efektif.

Untuk mencapai tujuan dari penelitian ini, penulis akan memanfaatkan sumber data resmi dari Susenas (Survei Sosial Ekonomi Nasional) dan BPS (Badan Pusat Statistik Indonesia, 2024). Sebagai tambahan, penulis juga memberikan dasar teoritis yang kuat dan mendukung argumen dengan mengacu pada literatur dan sumber informasi lainnya yang relevan dalam penelitian ini.

2. KAJIAN PUSTAKA

A. Data Mining

Data mining ialah suatu tahapan dalam mengekstraksi informasi yang bermanfaat dari sebuah data besar (*Big data*). Hal ini melibatkan penerapan pendekatan statistik, algoritma, dan prosedur untuk mengidentifikasi pola dan korelasi dalam data (Sari dkk., 2018). Penerapan *data mining* tersebar di berbagai sektor, termasuk bisnis, ilmu pengetahuan, dan teknologi informasi, dengan tujuan mengidentifikasi pola dalam data yang mendukung pengambilan keputusan yang lebih baik. Serta, bertujuan untuk mengidentifikasi sebuah informasi yang bermanfaat dan di ubah menjadi suatu format yang dapat dipahami dan dimengerti oleh pengguna.

B. Algoritma K-Means Clustering

Algoritma *K-Means Clustering* adalah metode analisis data atau teknik *data mining* yang menggunakan model tanpa supervisi (*unsupervised*) dan merupakan salah satu model yang mengelompokkan data melalui sistem partisi (Feryanto dkk., 2020). Metode *K-Means Clustering* berusaha mengelompokkan data ke dalam beberapa kategori, di mana data dalam satu kategori memiliki karakteristik yang mirip antara satu sama lain dan berbeda dari data dalam kategori lainnya (Fina Nasari & Surya Darma, 2015).

Algoritma *K-Means Clustering* adalah teknik non-hierarkis yang pada awalnya memilih sejumlah titik data untuk menjadi pusat awal dari *cluster*. *K-Means Clustering* dapat mengelompokkan data dalam jumlah besar dengan waktu pemrosesan yang cukup cepat dan efisien. Namun, metode ini juga memiliki kelemahan, khususnya terkait dengan pemilihan pusat awal *cluster* (Butsianto & Mayangwulan, 2020). Hasil yang akan dihasilkan oleh algoritma *K-Means Clustering* sangat dipengaruhi oleh pemilihan nilai awal *cluster*. Berikut adalah penjelasan dan langkah-langkah dalam menjalankan algoritma *K-Means Clustering* (Gustientiedina dkk., 2019):

1. Menetapkan k sebagai jumlah *cluster* yang akan dibentuk.
2. Menentukan k titik pusat awal *cluster* secara acak. Penentuan awal *centroid* dilakukan secara acak dari data objek yang sudah tersedia sebanyak k *cluster*.
3. Menghitung jarak dari setiap data objek ke masing-masing *centroid* dari beberapa *cluster* yang ada menggunakan metode perhitungan jarak *Euclidean Distance* yang dirumuskan sebagai berikut:

$$D(i, j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2} \quad (1)$$

Keterangan:

$D(i, j)$ = Merupakan jarak data ke i untuk ke pusat *cluster* j

x_{ki} = Pada data ke i atribut data ke k

x_{kj} = Pada titik pusat ke j untuk atribut ke k

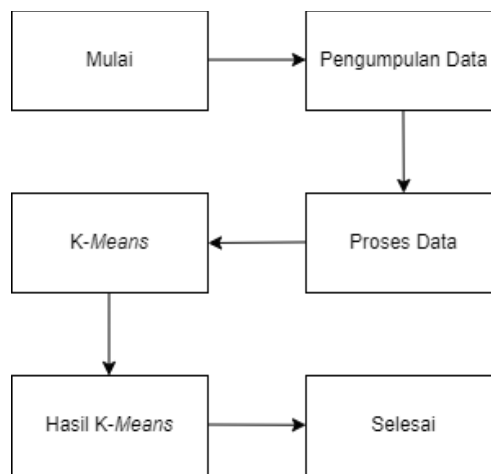
4. Mengalokasikan masing-masing data objek ke dalam *cluster* dengan mengukur jarak kedekatan objek terhadap titik pusat *cluster*.
5. Melakukan iterasi dan menentukan posisi *centroid* baru menggunakan persamaan. Pusat *cluster* yang baru adalah nilai rata-rata dari semua objek data yang ada dalam *cluster* tersebut.
6. Mengulangi proses perhitungan jika data objek masih berubah. Jika pusat *cluster* tidak lagi berubah, maka proses *Clustering* dengan algoritma *K-Means Clustering* dapat dikatakan selesai.

C. RapidMiner

RapidMiner menyediakan antarmuka *drag-and-drop* yang memudahkan pengguna dalam membuat alur kerja untuk pemrosesan dan analisis data. Platform ini mendukung berbagai sumber data, termasuk basis data, file datar, serta platform big data seperti *Hadoop* dan *Spark*. Selain itu, perangkat lunak ini dilengkapi dengan berbagai operator bawaan yang merupakan komponen utama dari alur kerja, mencakup semua tahap dalam proses penambangan data, termasuk pembersihan data, pemilihan fitur, dan pemodelan (Nahjan dkk., 2023).

3. METODE PENELITIAN

Langkah-langkah dalam penelitian ini dimulai dengan pengumpulan data, diikuti oleh pemrosesan data, penerapan *K-Means Clustering* menggunakan *RapidMiner*, dan diakhiri dengan analisis hasil dari proses *K-Means Clustering*.



Gambar 1. Diagram alir penelitian

A. Jenis Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berbentuk deret waktu (*time series*) yang bersifat kuantitatif, berupa data tahunan dalam bentuk angka dari tahun 2021 hingga 2023 (selama 3 tahun). Data sekunder adalah data primer (data asli) yang telah dikumpulkan oleh pihak lain sebelumnya dan akan diolah kembali untuk kebutuhan penelitian yaitu data Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan dan Provinsi.

B. Sumber Data

Data dalam penelitian ini menggunakan data yang diperoleh dari Badan Pusat Statistik (BPS) Indonesia, mencakup periode tiga tahun dari 2021 hingga 2023. Penelitian ini memanfaatkan data mengenai tingkat penyelesaian pendidikan menurut jenjang pendidikan yaitu SD, SMP, dan SMA.

C. Implementasi Algoritma *K-Means Clustering*

1. Penetapan Jumlah *Centroid*

Penetapan jumlah *centroid* dalam penelitian ini dilakukan dengan dipilih secara *random* sebanyak 3 *centroid* yang ditandai dengan warna kuning (Kep.Riau (10), Nusa Tenggara Timur (19), dan Gorontalo (29)).

Tabel 1. *Centroid* awal

Jenjang	SD / Sederajat			SMP / Sederajat			SMA / Sederajat		
Data ke	2021	2022	2023	2021	2022	2023	2021	2022	2023
10	99,26	98,58	98,66	95	95,4	95,85	84,98	87,71	88,1
19	94,29	95,13	95,33	79,65	81,82	81,56	54,27	58,4	55,58

Jenjang	SD / Sederajat			SMP / Sederajat			SMA / Sederajat		
Data ke	2021	2022	2023	2021	2022	2023	2021	2022	2023
29	95,93	97,15	95,13	86,09	84,14	84,04	56,22	55,18	54,79

2. Menghitung jarak dari setiap data objek ke masing-masing *centroid* dari beberapa *cluster* yang ada menggunakan metode perhitungan jarak *Euclidean Distance*.

Tabel 2. Menghitung Jarak Setiap Data SD ke Centroid

34 Provinsi	SD / Sederajat			Clusteriasi jenjang SD			Jarak kedekatan	
	2021	2022	2023	C1	C2	C3	Minimum	Cluster
Aceh	98,57	98,74	98,75	0,714002801	6,561013641	4,754166594	0,7140028	1
Sumatera Utara	95,29	97,87	95,81	4,938370176	2,956010825	1,179152238	1,1791522	3
Sumatera Barat	96,91	98,2	98,09	2,447815353	4,889468274	3,29006079	2,4478154	1
Riau	98,54	97,62	97,76	1,5	5,492494879	3,734956492	1,5	1
Jambi	97,82	97,53	97,58	2,083866598	4,825287556	3,117531074	2,0838666	1
Sumatera Selatan	98,16	97,65	97,1	2,123322868	4,945725427	3,017250404	2,1233229	1
Bengkulu	98,33	98,41	98,67	0,94546285	6,183494158	4,458609649	0,9454628	1
Lampung	96,61	96,45	96,01	4,310672801	2,754487248	1,314077623	1,3140776	3
Kep. Bangka Belitung	98,16	98,38	97,92	1,340746061	5,678688229	3,777552117	1,3407461	1
Kep. Riau	99,26	98,58	98,66	0	6,905961193	5,059120477	0	1
Dki Jakarta	98,45	99,08	99,09	1,044509454	6,858986806	5,075125614	1,0445095	1
Jawa Barat	98,06	98,01	98,42	1,35	5,661748846	4,012555296	1,35	1
Jawa Tengah	98,48	98,91	98,95	0,895209473	6,704394081	4,918587196	0,8952095	1
Di Yogyakarta	97,76	98,71	98,78	1,510397299	6,062986063	4,370926675	1,5103973	1
Jawa Timur	98,82	97,24	97,15	2,066228448	5,318402016	3,527123474	2,0662284	1
Banten	97,02	97,55	98,43	2,476166392	4,787410574	3,498299587	2,4761664	1
Bali	98,71	98,47	98,11	0,785557127	6,198419153	4,283830062	0,7855571	1
Nusa Tenggara Barat	91,84	92,35	93,41	11,01960979	4,173403887	6,536551078	4,1734039	2
Nusa Tenggara Timur	94,29	95,13	95,33	6,905961193	0	2,60959767	0	2
Kalimantan Barat	97,45	98,51	97,47	2,167279401	5,097999608	3,104126286	2,1672794	1
Kalimantan Tengah	95,67	94,68	95,99	5,935233778	1,594521872	2,628326464	1,5945219	2
Kalimantan Selatan	96,82	99,18	97,88	2,630969403	5,413492403	3,532067383	2,6309694	1
Kalimantan Timur	95,77	96,94	96,41	4,464549249	2,57544171	1,306942998	1,306943	3
Kalimantan Utara	96,1	96,74	96,18	4,418325475	2,567235868	1,13995614	1,1399561	3
Sulawesi Utara	96,19	97,19	97,56	3,544996474	3,581410337	2,44419721	2,4441972	3
Sulawesi Tengah	97,3	98,05	98,37	2,050999756	5,179584925	3,631046681	2,0509998	1
Sulawesi Selatan	95,58	97,24	97,83	4,003361088	3,51656082	2,724077826	2,7240778	3
Sulawesi Tenggara	93,44	95,12	93,69	8,399101142	1,847214119	3,52059654	1,8472141	2
Gorontalo	95,93	97,15	95,13	5,059120477	2,60959767	0	0	3
Sulawesi Barat	98,5	98,98	98,69	0,859360227	6,620891179	4,756826673	0,8593602	1
Maluku	96,97	97,72	98,3	2,472508847	4,76564791	3,384582692	2,4725088	1
Maluku Utara	91,81	93,94	92,69	10,61475388	3,812623768	5,764728962	3,8126238	2
Papua Barat	78,43	81,99	80,09	32,46477938	25,62141292	27,60936798	25,621413	2
Papua	97,37	97,82	97,83	2,199681795	4,792963593	3,132491022	2,1996818	1

3. Melakukan iterasi dan menentukan posisi *centroid* baru menggunakan persamaan. Pusat *centroid* yang baru adalah nilai rata-rata dari semua objek data yang ada dalam *cluster* tersebut.

Tabel 3. Iterasi ke 1

PEMETAAN JENJANG SD			PEMETAAN JENJANG SMP			PEMETAAN JENJANG SMA		
C1	C2	C3	C1	C2	C3	C1	C2	C3
21	6	7	13	4	17	7	8	19

Ulangi perhitungan sehingga hasil dari rekapitulasi didapatkan titik pusat baru (*centroid* baru). Setelah ditemukan titik pusat yang baru (*centroid* baru) tiap *cluster*, ulang Kembali dari tahap sebelumnya sampai titik pusat *cluster* tidak berganti lagi dan datanya tidak lagi berpindah dari satu *cluster* ke *cluster* lain.

Dikarenakan pada iterasi ke 2 data masih ada yang berubah jadi perhitungan masih akan terus berlanjut sampai tidak ada data yang berubah lagi. Selanjutnya menentukan *centroid* baru iterasi 3. Untuk perhitungan berhenti pada iterasi ke-9. Hasil perhitungan untuk iterasi ke-3 hingga iterasi ke-9 telah disertakan dalam lampiran. Hasil *cluster* yang didapatkan dari tiap data pada iterasi ke 9 tersebut terdapat pada tabel di bawah.

Tabel 4. Hasil Pengelompokan Data SD

Nama Cluster	Data SD	Jumlah Data SD
Cluster 1	1, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, 22, 25, 26, 27, 30, 31, 34	23
Cluster 2	33	1
Cluster 3	2, 8, 18, 19, 21, 23, 24, 28, 29, 32	10

Tabel 5. Hasil Pengelompokan Data SMP

Nama Cluster	Data SMP	Jumlah Data SMP
Cluster 1	1, 9, 10, 13, 15, 16, 17, 22, 24, 30, 31	11
Cluster 2	18, 19, 28, 33	4
Cluster 3	2, 3, 4, 5, 6, 7, 8, 11, 12, 14, 20, 21, 23, 25, 26, 27, 29, 32, 34	19

Tabel 1. Hasil Pengelompokan Data SMA

Nama Cluster	Data SMA	Jumlah Data SMA
Cluster 1	1, 9, 10, 13, 16, 22	6
Cluster 2	2, 3, 4, 5, 6, 7, 8, 11, 14, 15, 17, 20, 21, 24, 26, 27, 30, 31, 34	19
Cluster 3	12, 18, 19, 23, 25, 28, 29, 32, 33	9

Berdasarkan temuan *cluster* yang terbentuk pada tabel di atas, berikut adalah *range* nilai untuk *cluster* yang terbentuk berdasarkan hasil analisis:

Range nilai untuk C1: (97,97 - 99,26) **SD**, (92,66 – 95) **SMP**, (78,17 - 88,1) **SMA**.

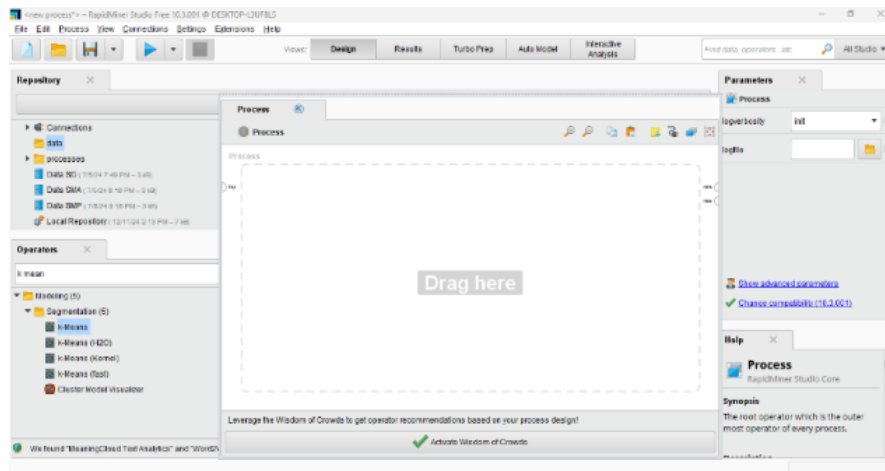
Range nilai untuk C2: (90,91 - 95,13) **SD**, (76,44 - 83,25) **SMP**, (54,27 - 66,25) **SMA**.

Range nilai untuk C3: (95,13 - 97,92) **SD**, (84,04 - 95,85) **SMP**, (53,46 - 88,1) **SMA**.

4. HASIL DAN PEMBAHASAN

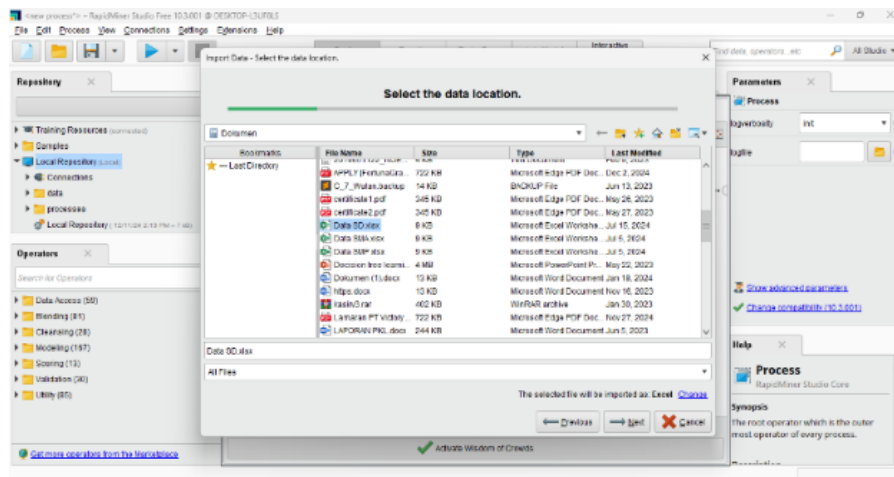
A. Perhitungan dan Implementasi Menggunakan *Tools RapidMiner*.

Pengolahan data ini diuji dengan menggunakan *tools RapidMiner* untuk mengetahui hasil dari 3 *cluster*, data yang akan diuji disimpan dalam format .xlsx (*excel*). Sehingga dapat dianalisa dengan *RapidMiner*, berikut ini adalah tampilan awal dari *RapidMiner* pada gambar 2 dibawah.



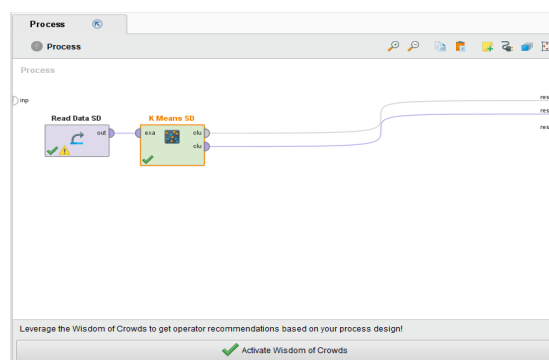
Gambar 2. Tampilan awal *RapidMiner Studio*

Proses selanjutnya merupakan *import* data pada bagian Repository, seperti pada gambar 3 di bawah. Kemudian pilih data yang akan digunakan untuk pengujian yaitu menggunakan 3 data jenjang SD, SMP, dan SMA.



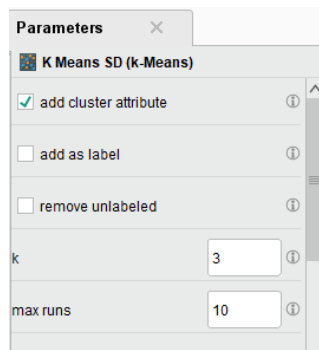
Gambar 3. Open File Excel Data SD

Pengujian pertama menggunakan Data SD, Langkah berikutnya adalah memilih operator yang akan digunakan, kemudian melakukan *drag and drop* operator seperti yang ditunjukkan pada Gambar 4 di bawah ini. Selanjutnya, seret data SD dari *Local Repository*. Cari operator K-Means pada bagian *operators*, lalu tarik ke panel proses dan sambungkan.



Gambar 4. Operator data dan clustering data SD

Pemodelan *clustering* dengan algoritma *K-Means* dengan jumlah *cluster* sebanyak 3 buah maka dengan menggunakan *Rapidminer*, *cluster* yang terbentuk juga sebanyak 3, tampilan *parameters* di bagian kolom K berfungsi untuk menguji *cluster* yang akan dibentuk, dilampirkan pada gambar 5 di bawah ini:



Gambar 5. Parameter Clustering Data SD

Cluster Model

Cluster 0: 28 items
Cluster 1: 1 items
Cluster 2: 5 items
Total number of items: 34

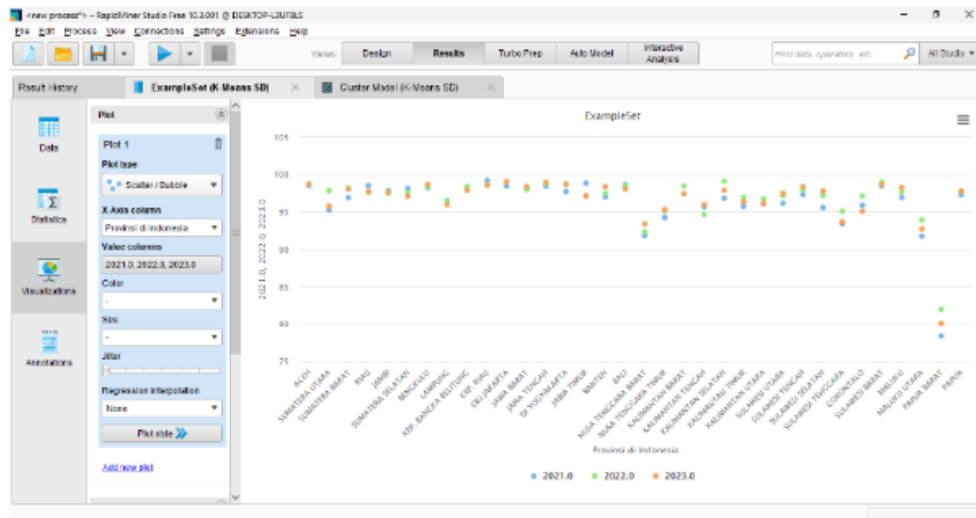
Gambar 6. Hasil Cluster Data SD

Setelah melakukan *run* dengan ini muncul *cluster* model pada data Tingkat penyelesaian jenjang SD, dapat terlihat jarak *centroid* antar *cluster*. Hasil *cluster* pada *Rapidminer* diawali dengan *cluster* 0 (nol), tetapi untuk teori *clustering* diawali menggunakan *cluster* 1. Jadi dalam *cluster* 0 pada *Rapidminer* sama dengan *cluster* 1 (C1) pada perhitungan skema. Hasil yang diperoleh dari perhitungan menggunakan *Rapidminer* seperti dalam Tabel 1.

Row No.	id	cluster	Provinsi BIL.	2021.0	2022.0	2023.0
1	1	cluster_0	ACEH	99.579	99.740	99.756
2	2	cluster_0	SUMATERA...	99.290	97.670	99.910
3	3	cluster_0	SUMATERA B.	99.919	98.200	99.999
4	4	cluster_0	RIAU	99.549	97.620	97.780
5	5	cluster_0	JAMBI	97.829	97.530	97.580
6	6	cluster_0	SUMATERA B.	99.189	97.650	97.190
7	7	cluster_0	BENGKULU	99.339	99.410	99.970
8	8	cluster_0	LAMPUNG	99.019	99.450	99.910
9	9	cluster_0	KEP. BANGK.	99.199	99.590	97.920
10	10	cluster_0	KEP. RIAU	99.269	99.550	99.940
11	11	cluster_0	DKI JAWART.	99.409	99.080	99.999
12	12	cluster_0	JAWA BARAT	99.099	99.010	99.420
13	13	cluster_0	JAWA TENGAH	99.489	99.010	99.960
14	14	cluster_0	DI YOGYAKA.	97.799	99.710	99.780
15	15	cluster_0	JAWA TIMUR	99.929	97.240	97.100

Gambar 7. Data dalam setiap cluster SD

Gambar diatas adalah data pengelompokan tingkat SD dari setiap jumlah *cluster* yang sudah terbentuk.

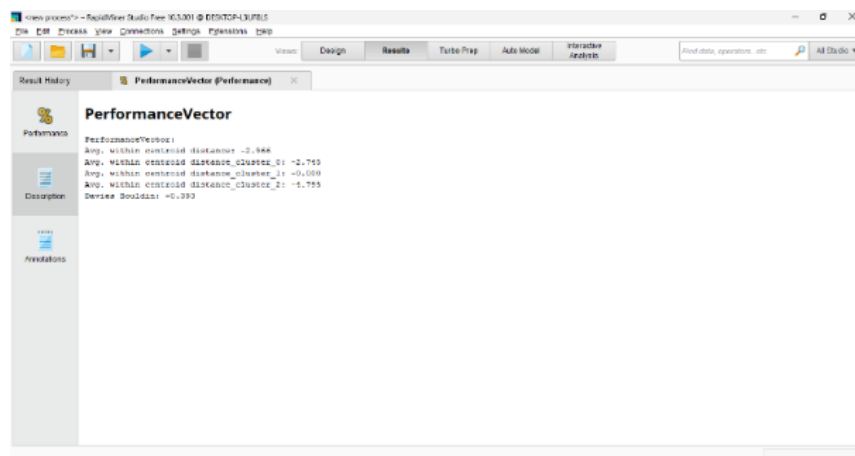


Gambar 8. Grafik RapidMiner data SD

Dibawah ini adalah tabel hasil pengelompokan data tingkat SD menggunakan *tools RapidMiner*.
Tabel 7. Hasil Pengelompokan Data SD Dengan RapidMiner

Nama Cluster	Data SD	Jumlah Data SD
Cluster 1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, 22, 23, 24, 25, 26, 27, 29, 30, 31, 34	28
Cluster 2	33	1
Cluster 3	18, 19, 21, 28, 32	5

Untuk mengevaluasi kinerja metode *K-means* dalam mengelompokkan data, dilakukan pengujian performa menggunakan *tools RapidMiner*. Pengujian ini memanfaatkan fitur *Cluster Distance Performance*, dan hasil yang diperoleh adalah sebagai berikut:



Gambar 4. 1 Hasil evaluasi cluster data SD

Avg. within centroid distance merupakan rata-rata jarak dalam *cluster* yang dihitung berdasarkan rata-rata jarak antara *centroid* dengan seluruh data dalam *cluster*. *Avg. within centroid distance* yang didapatkan adalah -2.966. Sedangkan untuk evaluasi *cluster* menggunakan metode *Davies-Bouldin Index*, apabila semakin kecil nilai DBI yang di dapat (non negatif ≥ 0), maka nilai *cluster* yang didapatkan dari pengelompokan dengan algoritma *K-Means Clustering* semakin baik. Nilai *Davies-Bouldin Index* yang didapatkan adalah 0.393.

5. KESIMPULAN

Berdasarkan hasil uraian permasalahan dan pengujian pada bab yang telah dilakukan sebelumnya dapat disimpulkan sebagai berikut ini:

- Berdasarkan hasil penelitian, ditemukan bahwa tingkat penyelesaian pendidikan di jenjang SD menunjukkan *cluster* tertinggi berada pada *cluster* pertama dengan jumlah 23 data, sedangkan hasil perhitungan menggunakan *RapidMiner* mengelompokkan 28 data dengan nilai *Davies-Bouldin Index* (DBI) sebesar 0.393, yang mencerminkan kualitas klasterisasi yang cukup baik.
- Untuk jenjang SMP, *cluster* tertinggi berada pada *cluster* ketiga dengan 19 data, sementara *RapidMiner* mengelompokkan sebanyak 26 data dengan nilai DBI 0.489.
- Pada jenjang SMA menunjukkan *cluster* tertinggi berada pada *cluster* kedua dengan 19 data, sedangkan *RapidMiner* menghasilkan 25 data dengan nilai DBI sebesar 0.562.

Hasil penelitian ini menunjukkan bahwa tingkat penyelesaian pendidikan bervariasi di antara provinsi-provinsi di Indonesia, dengan pola pengelompokan yang berbeda di setiap jenjang pendidikan. Nilai DBI yang diperoleh menunjukkan efektivitas metode *K-Means Clustering* dalam mengelompokkan data pendidikan.

Penelitian ini memiliki beberapa keterbatasan, antara lain menggunakan data sekunder yang telah tersedia secara *online*, jumlah sampel yang dianalisis hanya sebanyak 34 data, dan jumlah *cluster* yang digunakan terbatas yakni hanya 3 *cluster*. Untuk penelitian mendatang, disarankan menambahkan parameter atau variabel lain dalam metode *clustering*, serta meningkatkan jumlah sampel agar hasilnya lebih representatif. Selain itu, penelitian lanjutan dapat dilakukan dengan membandingkan metode yang digunakan dengan *Fuzzy C-Means* untuk mengelompokkan atau klasifikasi data.

6. DAFTAR PUSTAKA

- Amri, M. A., Windarto, A. P., Wanto, A., & Damanik, I. S. (2019). Analisis Metode K-Means Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas Yang Dimiliki Desa. *Komik (konferensi nasional teknologi informasi dan komputer)*, 3(1), 674–679.
- Badan Pusat Statistik Indonesia, 2024. *Tingkat Penyelesaian Pendidikan Menurut Jenjang Pendidikan Dan Provinsi, 2021-2023*. <https://www.bps.go.id/id/statistics-table/2/mtk4mcm/tingkat-penyelesaian-pendidikan-menurut-jenjang-pendidikan-dan-provinsi.html>.
- Butsianto, S., & Mayangwulan, N. T. (2020). Penerapan Data Mining Untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, 3(3), 187–201.
- Dewi, S. M., Windarto, A. P., Damanik, I. S., & Satria, H. (2019). Analisa Metode K-Means Pada Pengelompokan Kriminalitas Menurut Wilayah. *Seminar Nasional Sains & Teknologi Informasi (Sensasi)*, 620–625.
- Direktorat, S., Statistik, I., Analisis, D., & Statistik, P. (2008). *Profil Indikator Pendidikan Di Indonesia (Laporan Sosial Indonesia 2009)*. <http://andri-mustawi.blogspot.com>
- Feryanto, Kesuma, F. T., & Tamba, S. P. (2020). Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering. *Jurnal sistem informasi dan ilmu komputer prima(jusikom prima)*, 2(2), 67–72. <https://doi.org/10.34012/jusikom.v2i2.376>
- Fina Nasari, & Surya Darma, S. (2015). Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru. *Seminar Nasional Teknologi Informasi Dan Multimedia 2015*, 73–78.
- Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi Dan Sistem Informasi*, 5(1), 17–24.
- Indriyani, F., & Irfiani, E. (2019). Clustering Data Penjualan Pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means. *Juita : Jurnal Informatika*, 7(2), 109.
- Nabila, Z., Isnain, A. R., Permata, & Abidin, Z. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (Jtsi)*, 2(2), 100.

- Nahjan, M. R., Heryana, N., & Voutama, A. (2023). Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell. Dalam *Jurnal Mahasiswa Teknik Informatika*, 7(1).
- Sari, R. W., Wanto, A., & Windarto, A. P. (2018). Implementasi Rapidminer Dengan Metode K-Means (Study Kasus: Imunisasi Campak Pada Balita Berdasarkan Provinsi). *Komik (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 2(1), 224–230.