

# Data Mining Untuk Klasterisasi Provinsi Di Indonesia Berdasarkan Data Pokok Pendidikan Nasional

Rossa Amelia Manik<sup>1</sup>, Atik Ariesta<sup>2</sup>

<sup>1</sup>Fakultas Teknologi Informasi, Sistem Informasi, Universitas Budi Luhur, Jakarta, Indonesia

<sup>2</sup>Fakultas Teknologi Informasi, Manajemen Informatika, Universitas Budi Luhur, Jakarta Indonesia

E-mail: <sup>1</sup>rossaamelia037@gmail.com, <sup>2</sup>atik.ariesta@budiluhur.ac.id

(\*: corresponding author)

**Abstrak**—Tidak semua wilayah memiliki pemerataan dalam pendidikannya, oleh karena itu diperlukan adanya perhatian khusus oleh pemerintah terhadap wilayah yang dinilai kurang dalam pemerataan pendidikannya. Pengelompokan serta penggolongan tingkatan Pendidikan suatu wilayah akan mempermudah pemerintah dalam pertimbangan wilayah mana saja yang memerlukan atensi lebih dalam penyelenggaraan Pendidikan. Dengan adanya pengelompokan tersebut akan dapan dilihat wilayah mana saja yang jumlah pendidikannya terbilang rendah atau tinggi pada suatu wilayah. Data yang diperoleh dengan kurun waktu semester Genap 2021- 2022. Pada penelitian ini Teknik dari data mining untuk klasterisasi Provinsi Di Indonesia yaitu dengan menggunakan algoritma K-Means. Data dari penelitian ini bersumber dari website dataset [dapo.kemendikbud.go.id](http://dapo.kemendikbud.go.id). Dataset terdiri dari 34 provinsi data. Yang terdiri dari total sekolah, total guru, total peserta didik, dan total tendik pada kategori SD, SMP, SMA, dan SMK. Data yang telah dikumpulkan akan diolah dengan melakukan clustering dalam 2 cluster yaitu Cluster 0 (Tingkat Pendidikan Tinggi) terdapat 3 provinsi di Indonesia yang berada pada wilayah Indonesia Bagian Barat dan Pada Cluster 1 (Tingkat Pendidikan Rendah) terdapat 31 Provinsi. Hasil yang diperoleh dapat menjadi masukan bagi pemerintah dan dinas sosial dalam menganalisa provinsi menurut tingkat pendidikan.

**Kata kunci:** *Data Mining, Pendidikan, Algoritma Clustering, K-Means*

**Abstract**—Not all regions have equity in education, therefore special attention is needed by the government for areas that are considered lacking in educational equity. Grouping and classifying the level of education in a region will make it easier for the government to assess which areas require more attention in implementing education. With this grouping, it will be seen which areas have relatively low or high education levels in a region. Data obtained over the Even semester 2021- 2022. In this study, the data mining technique for clustering the Provinces of Indonesia is by using the K-Means algorithm. The data from this study were sourced from the dataset website [Dapo.kemendikbud.go.id](http://Dapo.kemendikbud.go.id). The dataset consists of 34 data provinces. Which consists of total schools, teachers, students, and students in the categories of SD, SMP, SMA, and SMK. The data that has been collected will be processed by clustering in 2 clusters, namely Cluster 0 (Higher Education Level) there are 3 provinces in Indonesia which are in the western part of Indonesia, and Cluster 1 (lower education level) there are 31 provinces. The results obtained can be used as

*input for the government and social services in analyzing provinces by the level of education.*

**Keywords:** *Data Mining, Education, Algorithm Clustering, K-Means*

## I. PENDAHULUAN

Pendidikan adalah kebutuhan oleh tiap manusia dalam menjalani kehidupan dan mempersiapkan kualitas Sumber Daya Manusia pada suatu Negara atau bangsa [1]. Sistem pendidikan Indonesia terdiri dari empat tingkatan: Sekolah Dasar (SD), Sekolah Menengah Pertama (SMP), Sekolah Menengah Atas (SMA) dan Sekolah Menengah Kejuruan (SMK). Pada empat jenjang tersebut, sekolah negeri saat ini mendominasi sistem pendidikan Indonesia dengan porsi 52 persen. Selain itu, 48 persen dimiliki oleh swasta [2]

Dalam penyelenggaraan Pendidikan, penyediaan sekolah dan tenaga pendidik merupakan salah satu factor pendukung dalam kegiatan belajar mengajar yang efektif dan efisien[3] Tidak semua wilayah memiliki pemerataan dalam pendidikannya, oleh karena itu diperlukan adanya perhatian khusus oleh pemerintah terhadap wilayah yang dinilai kurang dalam pemerataan pendidikannya. Pengelompokan serta penggolongan tingkatan Pendidikan suatu wilayah akan mempermudah pemerintah dalam pertimbangan wilayah mana saja yang memerlukan atensi lebih dalam penyelenggaraan Pendidikan. Dengan adanya pengelompokan tersebut akan dapan dilihat wilayah mana saja yang jumlah pendidikannya terbilang rendah atau tinggi pada suatu wilayah. Dari perihal tersebut dapat dijadikan sebagai bahan pertimbangan pemerintah, wilayah mana saja yang memerlukan atensi lebih dalam penyelenggaraan Pendidikan sehingga tidak tertinggal oleh wilayah lainnya.

Terdapat metode – metode dari penelitian terdahulu yang melakukan klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik [4], Pengelompokan Wilayah Berdasarkan Faktor Pendukung Pendidikan Dengan Jumlah Sekolah dan Jumlah Guru Menggunakan Algoritma K-Means[5], Analisis Metode K-Means Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas Yang Dimiliki Desa [6],

Pemetaan dan Klasterisasi Sekolah Muhammadiyah di Kabupaten Berdasarkan Fasilitas, Pendidik dan Tenaga Pendidik Menggunakan *K-Means Clustering*[7], dan penelitian lain dengan menggunakan algoritma berbeda yaitu Implementasi Metode *K-Medoids Clustering* untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan. [8].

Data Mining atau penambangan data merupakan proses pencarian pola yang menarik dan tersembunyi dari kumpulan data yang berukuran besar. Data tersebut tersimpan di basis data, Gudang data, atau tempat penyimpanan data lainnya. Penambangan data merupakan salah satu bagian dari kecerdasan buatan. *Clustering* atau klasterisasi adalah metode pengelompokan data dalam proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*. objek yang di dalam cluster memiliki karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering* [9].

## II. TINJAUAN PUSTAKA

### A. Data Mining

*Data mining* adalah proses menemukan dan menganalisis data dalam jumlah besar dengan tujuan menemukan pola atau informasi yang menarik dari sejumlah besar data yang tersimpan dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma yang tepat sangat bergantung pada tujuan dan proses *knowledge discovery in database* (KDD) secara keseluruhan[10]

### B. CRISP-DM (Cross Standard Industry for Data Mining)

*CRISP-DM* merupakan metode yang banyak digunakan oleh para ahli dengan menggunakan proses pemodelan data di dalamnya. Tujuan dari metode *CRISPDM* yaitu untuk menemukan pola yang menarik dan memiliki makna pada data yang digunakan. *CRISP-DM* memiliki tahapan dan kerangka kerja yang terstruktur sehingga pengguna metode ini akan lebih terarah dan mengetahui langkah yang harus dikerjakan dalam Penelitian [11].

### C. Algoritma K-means

*K-Means clustering* ialah metode analisis cluster yang bertujuan untuk memecah objek menjadi k cluster kemudian diamati di mana setiap objek cluster diperoleh melalui rata-rata terdekat. Nilai rata-rata pada cluster kemudian dihitung secara berulang pada proses awal. Adapun tahap-tahap melakukan *K-Means Clustering* sebagai berikut [12]

#### 1) Tahapan Proses Pada *K-Means*

Berikut adalah tahapan-tahapan untuk menggunakan Algoritma *K-Means*:

a. Menentukan jumlah *cluster* (k) yang diinginkan pada dataset

b. Menentukan centroid secara random pada tahap awal

$$\bar{V}_j = \frac{1}{N_i} \sum_{k=0}^{N_1} X_{kj}, \quad (2.1)$$

di mana:

$V_{ij}$  = *centroid*/rata-rata *cluster* ke-i untuk variabel ke-j

$N_i$  = jumlah data yang menjadi anggota *cluster* ke-i

i, k = indeks dari *cluster*

j = indeks dari variabel

$X_{kj}$  = nilai data ke-k yang ada di dalam cluster tersebut untuk variabel ke-j

c. Menentukan jarak terdekat setiap data dengan *centroid*. *Euclidean distance* digunakan untuk menghitung jarak terdekat dengan *centroid*.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}, \quad (2.2)$$

De adalah *Euclidean*

*Distance* i adalah banyaknya objek

(x,y) merupakan koordinat objek (s, t) merupakan koordinat *centroid*

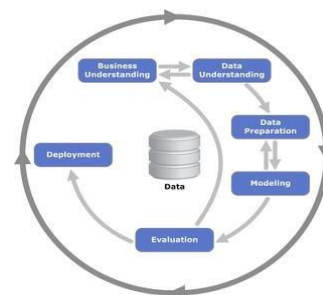
d. Menghitung kembali custer dengan keanggotaan cluster baru. Rata-rata dari semua data dijadikan sebagai pusat cluster. Proses berhenti 20 endid nilai pada *cluster* tidak berubah lagi.

### D. Klasterisasi

Klasterisasi merupakan salah satu metode untuk mengidentifikasi kelas berdasarkan obyek-obyek. Reprsentasi kesamaan berbeda – beda pada model klasterisasi satu dengan yang lain. Di dalam banyak model, konsep kesamaan meliputi Manhattan Distance, Euclidean Distance dan lain – lain. Tetapi fungsi jarak tidak selalu cukup untuk mengambil korelasi – korelasi antara obyek – obyek yang jauh[10]

## III. METODE PENELITIAN

Pada tahapan ini menerapkan metode *CRISP-DM* (*Cross Industry Standard Process Model for Data Mining*) dengan beberapa tahapan didalamnya. Model Proses *CRISP-DM* dapat dijelaskan sebagai berikut:



Gambar 1. Proses *CRISP-DM*

Berdasarkan Gambar 1, berikut adalah penjelasan tentang tahapan-tahapan yang dilakukan:

#### a. *Business Understanding*

Tahap pertama subjek pada penelitian ini adalah seluruh provinsi di Indonesia. Oleh karena itu perlu adanya strategi

dengan clustering atau pengelompokan data seluruh provinsi di Indonesia.

b. *Data Understanding*

Peneliti melakukan pemahaman terhadap kebutuhan data terkait dengan korelasi Pendidikan pada seluruh Provinsi di Indonesia. Penelitian ini menggunakan data dari Data Pokok Pendidikan Nasional kurun waktu semester Genap 2021-2022. Data Pendidikan meliputi 34 Provinsi di Indonesia yang meliputi data: Jumlah Peserta Didik, Jumlah Guru, Jumlah Tendik, dan Jumlah Sekolah. Data tersebut didapatkan dari total seluruh data pada jenjang TK, SD, SMP, SMA, dan SMK pada setiap provinsi.

c. *Data Preparation*

Pada tahap ini adalah tahapan dari semua kegiatan untuk membangun dataset akhir dari data mentah. Tahapan ini melakukan proses data Transformasi dan data Normalisasi. Setelah itu, data digabung ke dalam format yang sesuai untuk diproses dalam data mining.

d. *Modeling*

Pada tahap pemodelan dilakukan pemilihan dan penerapan teknik pemodelan serta beberapa parameternya yang akan disesuaikan untuk mendapatkan cluster yang optimal. Terdapat beberapa pemodelan untuk menyelesaikan masalah yang sama pada data mining. Dalam penelitian ini akan menerapkan algoritma *K-Means* dan menggunakan tool Rapid Miner.

e. *Evaluation*

Di tahap ini akan melakukan evaluasi memakai metode Davies Bouldin Index untuk memilih jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster*.

f. *Deployment*

Tahapan ini dilakukan setelah menyelesaikan tahapan-tahapan diatas untuk menghasilkan laporan klasterisasi dalam menentukan jumlah tingkat Pendidikan yang tinggi dan rendah pada setiap Provinsi di Indonesia.

- 2) Data Peserta adalah hasil akumulasi dari jumlah peserta didik TK, SD, SMP, SMA, dan SMK baik Laki – Laki maupun Perempuan untuk Kabupaten/Kota untuk setiap Provinsi.
- 3) Data Guru adalah hasil akumulasi dari jumlah Guru TK, SD, SMP, SMA, dan SMK baik Laki – Laki maupun Perempuan untuk Kabupaten/Kota di setiap Provinsi.
- 4) Data Tendik adalah hasil akumulasi dari jumlah Tenaga Didik/Pegawai pada TK, SD, SMP, SMA, dan SMK baik laki – laki maupun Perempuan untuk setiap Provinsi.

Berikut merupakan data hasil akumulasi semua atribut data berdasarkan jumlah pada seluruh Kabupaten/Kota pada masing – masing Provinsi. Salah satu Contoh Data Pada Provinsi Aceh.

Wilayah	Sekolah				Peserta Didik				Guru				Tendik			
	SD	SMP	SMA	SMK	SD	SMP	SMA	SMK	SD	SMP	SMA	SMK	SD	SMP	SMA	SMK
Kab. Aceh Utara	366	143	54	27	59.478	24.766	13.804	4.886	5.239	2.757	1.465	684	844	920	479	240
Kab. Bireuen	233	100	37	12	44.286	15.115	7.724	3.388	3.621	2.166	1.157	460	541	621	387	126
Kab. Aceh Timur	290	84	31	17	33.205	19.154	9.774	3.716	3.766	2.054	1.088	391	769	428	261	116
Kab. Aceh Besar	214	80	48	11	34.078	14.562	9.330	3.274	3.564	1.562	815	487	562	434	240	147
Kab. Aceh Selatan	205	60	33	12	30.324	15.053	8.117	3.145	2.475	1.319	971	445	584	514	245	86
Kab. Pidie	279	72	31	14	28.952	12.726	8.898	3.480	2.447	989	805	282	370	369	369	133
Kab. Aceh Tenggara	180	77	29	17	23.954	10.919	9.428	4.626	2.180	1.008	474	328	355	293	262	166
Kab. Aceh Tengah	201	52	20	6	24.594	10.288	6.451	2.578	2.071	960	500	254	392	224	244	132
Kab. Aceh Barat	155	61	22	11	19.927	8.445	6.382	1.489	1.845	944	737	343	401	245	122	63
Kab. Aceh Tamiang	170	59	27	10	19.697	6.106	4.105	2.464	1.330	773	823	407	284	258	149	193
Kab. Bener Meriah	132	58	23	6	17.666	7.653	3.576	4.389	1.688	958	387	278	253	237	191	102
Kab. Aceh Jaya	98	37	14	7	15.879	6.245	4.592	2.739	1.748	830	472	223	311	253	118	43
Kab. Nagan Raya	137	41	18	5	14.072	6.531	3.750	3.524	1.322	723	385	489	237	204	137	84
Kab. Aceh Singkil	111	46	15	9	16.956	5.912	3.620	1.099	1.298	905	421	189	274	250	155	50
Kota Banda Aceh	91	34	29	9	15.127	5.631	3.834	858	1.515	658	418	97	231	204	107	111
Kab. Simeulue	115	46	26	8	16.716	5.150	3.020	1.308	1.416	645	449	147	227	165	120	44
Kab. Aceh Barat Daya	111	32	15	5	12.137	5.379	4.350	1.022	1.383	651	425	133	309	141	97	21
Kab. Pidie Jaya	89	32	14	7	12.259	4.714	2.934	1.442	1.040	595	321	359	264	122	66	28
Kota Lhokseumawe	75	35	12	13	12.141	5.254	3.440	1.165	1.244	580	291	165	241	148	104	34
Kota Subulussalam	86	35	14	6	8.971	3.806	2.194	832	969	471	240	114	182	105	69	51
Kota Langsa	67	25	10	10	9.805	4.005	3.779	967	931	349	247	157	114	81	58	36
Kab. Gayo Lues	92	42	16	2	11.190	3.573	2.720	309	931	484	259	61	133	133	83	18
Kota Sabang	25	9	3	1	3.787	1.588	1.182	359	372	215	102	40	85	59	25	14
Total	3.526	1.258	541	225	485.001	202.575	127.004	32.967	44.385	22.496	13.452	6.533	7.973	6.408	4.088	2.038

Gambar 1. Seluruh Atribut Data Sebelum di Akumulasi pada Provinsi Aceh

Sehingga seluruh atribut Data pada setiap Provinsi disatukan menjadi satu Dataset sebelum diolah seperti pada tabel 1 berikut ini:

TABEL I  
Data Hasil Akumulasi

Provinsi	Data Sekolah	Data Peserta Didik	Data Guru	Data Tendik
Prov. Aceh	5.550	867.547	86.866	20.507
Prov. Bali	3.178	755.926	45.104	15.418
Prov. Banten	7.576	2.048.123	98.827	25.489
Prov. Bengkulu	2.107	364.245	28.927	7.149
...	...	...	...	...
Prov. Sumatera	14.545	2.898.451	185.433	33.163

C. *Persiapan Data (Data Preparation)*

Di tahap ini data didapat sebanyak 34 *record* Provinsi di Indonesia pada semester Genap 2021 – 2022. Data *Preprocessing* pada penelitian ini dibagi menjadi beberapa langkah yaitu Transformasi data dan Normalisasi Data. Langkah – langkah yang dilakukan pada penelitian ini adalah sebagai berikut :

1) *Transformasi Data*

Pada tahapan ini untuk membedakan antara wilayah yang satu dengan wilayah yang lainnya dengan menambahkan Kode zona wilayah dan data wilayah pada setiap Provinsi di Indonesia. Berikut adalah Transformasi data yang dilakukan:

IV. PEMBAHASAN

A. *Pemahaman Bisnis (Business Understanding)*

Pemahaman bisnis dari penelitian ini ialah mengelompokkan dan menganalisis hasil cluster data pendidikan yang ada pada provinsi di Indonesia. Hasil pengelompokkan ini dapat memberikan saran kepada pengambil keputusan terhadap wilayah yang dinilai kurang dalam pemerataan Pendidikan.

B. *Pemahaman Data (Data Understanding)*

Penelitian ini menggunakan data dari Data Pokok Pendidikan Nasional yang diperoleh dari web <https://dapo.kemdikbud.go.id/> pada kurun waktu semester Genap 2021-2022. Data tersebut terdiri dari 34 provinsi di Indonesia yang terdiri dari data peserta didik, data guru, data tendik/ pegawai pada jenjang SD, SMP, SMA, SMK. Berikut data yang terdapat pada web tersebut:

- 1) Data Sekolah adalah hasil akumulasi dari jumlah sekolah TK, SD, SMP, SMA, dan SMK baik negeri maupun swasta untuk Kabupaten/Kota di setiap Provinsi.

TABEL II  
 Zona Wilayah

Nama Wilayah	Zona Wilayah
Indonesia Bagian Barat	1
Indonesia Bagian Tengah	2
Indonesia Bagian Timur	3

2) Normalisasi Data

Pada Tahapan ini untuk melakukan penskalaan nilai atribut dari data data sehingga bisa jatuh pada range tertentu. Pada penelitian ini menggunakan min – max normalization dengan rumus :

$$v' = \frac{v - \min A}{\max A - \min A} (new_{\max A} - new_{\min A}) + new_{\min A}$$

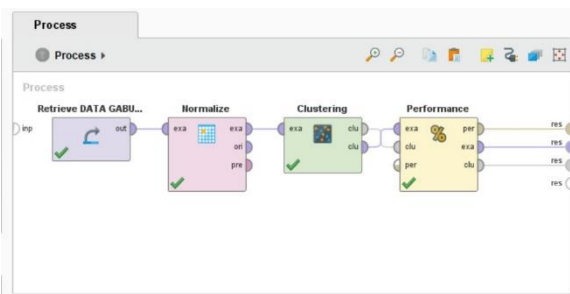
Berikut adalah tabel atribut data setelah melakukan proses normalisasi

TABEL III  
 Data Normalisasi

Provinsi	Normalisasi Data Sekolah	Normalisasi Data Peserta Didik	Normalisasi Data Guru	Normalisasi Data Tendik
Prov. Aceh	0,093	0,163	0,209	0,191
Prov. Bali	0,079	0,082	0,095	0,136
Prov. Banten	0,243	0,232	0,242	0,245
Prov. Bengkulu	0,029	0,045	0,051	0,047
...	...	...	...	...
Prov. Sumatera Selatan	0,182	0,214	0,252	0,241
Prov. Sumatera Utara	0,351	0,469	0,478	0,327

D. Modeling

Pada tahap ini peneliti melakukan tahap modeling untuk menentukan cluster yang optimal menggunakan K-Means. Berikut proses yang dilakukan dengan menggunakan operator Normalize, K-means, dan cluster distance performance dengan tools RapidMiner yang dapat dicermati di Gambar 5.



Gambar 2. Proses Algoritma K-Means

Setelah melakukan proses rapid miner, maka dapat dilihat hasil pada table 4 nilai DBI paling kecil dengan penentuan jumlah cluster terbaik adalah berjumlah 2 cluster dengan nilai 0,401. Semakin kecil nilai DBI maka semakin optimal cluster yang dihasilkan.

TABEL IV  
 Hasil Davies Bouldin Index

Penentuan Jumlah Cluster Terbaik	Hasil Davies Bouldin Index
2 Cluster	0.401
3 Cluster	0.469
4 Cluster	0.422

Selanjutnya, pada gambar 4 merupakan hasil cluster yang diperoleh

Cluster Model

```
Cluster 0: 3 items
Cluster 1: 31 items
Total number of items: 34
```

Gambar 3. Hasil Jumlah Cluster

Dengan menggunakan pemodelan K-means clustering seperti pada gambar . diatas, dengan jumlah data sebanyak 34 dan inialisasi jumlah cluster sebanyak 2 buah, cluster, sesuai dengan pendefinisian nilai k dengan jumlah cluster\_0 : 3 items, cluster\_1 : 31 items.

Selanjutnya, yaitu tabel hasil dataset menurut zona wilayah pada setiap cluster yang diolah terdapat pada table 5.

TABEL V  
 Dataset Zona Wilayah Pada Setiap Cluster

Cluster 0	Indonesia Bagian Barat: 3
	Indonesia Bagian Timur: 0
	Indonesia Bagian Tengah: 0
Cluster 1	Indonesia Bagian Barat: 13
	Indonesia Bagian Timur: 10
	Indonesia Bagian Tengah: 8

Selanjutnya terdapat nilai rata – rata pada seluruh Atribut Data Pada Cluster 0 yang terdapat pada gambar 5 berikut ini

	CLUSTER 0							
	PESERTA DIDIK		SEKOLAH		GURU		TENDIK	
	LAKI	PEREMPUAN	NEEGRI	SWASTA	LAKI	PEREMPUAN	LAKI	PEREMPUAN
SD	1.677.177	1.555.519	17.332	1.871	56.085	116.371	25.862	14.058
SMP	719.991	676.051	1.828	2.908	31.977	46.389	15.028	8.701
SMA	240.906	328.412	432	927	14.091	17.810	6.890	3.477
SMK	507.704	365.990	275	1.940	24.126	24.383	9.309	4.934

Gambar 4. Rata - Rata pada cluster 0

Selanjutnya terdapat nilai rata – rata pada seluruh Atribut Data Pada Cluster 0 yang terdapat pada gambar 6 berikut ini

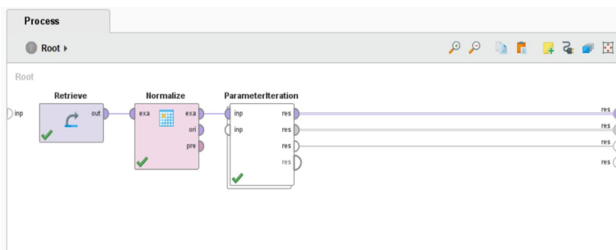
CLUSTER 1								
	PESERTA DIDIK		SEKOLAH		GURU		TENDIK	
	LAKI - LAKI	PEREMPUAN	NEGERI	SWASTA	LAKI - LAKI	PEREMPUAN	LAKI - LAKI	PEREMPUAN
SD	242.362	222.093	2.529	426	8.141	21.837	3.413	3.198
SMP	94.663	90.638	592	306	4.690	9.190	1.938	1.679
SMA	49.200	37.416	183	144	2.754	4.759	1.107	879
SMK	44.804	33.626	93	157	2.399	3.261	842	632

Gambar 5. Rata - Rata Pada cluster 1

Pada cluster 0 memiliki nilai rata – rata pada setiap atribut seperti jumlah peserta didik terbanyak Pada tingkat SD pada jenis kelamin laki – laki sebanyak 1.677.177 orang, jumlah sekolah terbanyak ada pada tingkat SD Negeri sebanyak 17.332 Sekolah, jumlah Guru terbanyak pada tingkat SD dengan jenis kelamin perempuan sebanyak 116.571 orang dan jumlah Tendik pada tingkat SD dengan jenis Kelamin Laki – laki sebanyak 25.862 orang. Selain itu, pada cluster 1 memiliki nilai rata – rata pada setiap atribut seperti jumlah peserta didik terbanyak Pada tingkat SD pada jenis kelamin laki – laki sebanyak 242.362 orang, jumlah sekolah terbanyak ada pada tingkat SD Negeri sebanyak 2.529 Sekolah, jumlah Guru terbanyak pada tingkat SD dengan jenis kelamin perempuan sebanyak 21.837 orang dan jumlah Tendik pada tingkat SD dengan jenis Kelamin Laki – laki sebanyak 3.413 orang.

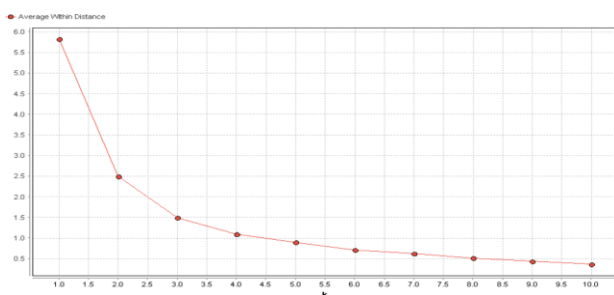
E. Evaluation

Pada bagian ini disampaikan mekanisme pengukuran performa model terbaik. Metode uji dari penelitian ini menggunakan *Elbow method*. Berikut proses dari pengujian yang menggunakan *tools Rapidminer*:



Gambar 6. Operator Pengujian

Proses diatas menggunakan operator loop parameters. Lalu dari hasil *processLog* menunjukkan nilai DB terkecil ialah 0,401 dengan jumlah 2 cluster yang optimal. Pada Gambar 10 berikut merupakan gambar yang menunjukan grafik Elbow



Gambar 7. Grafik Elbow

Gambar 10 adalah visualisasi kluster optimal dengan menggunakan metode Elbow. Berdasarkan gambar 10. diatas menunjukkan bahwa kluster dengan jumlah 2, 3 dan 4 memberikan sudut yang membentuk siku dan mengalami penurunan paling besar. Oleh karena itu dapat diperkirakan bahwa kluster optimal berjumlah 2, 3 atau 4 Cluster. Pada metode elbow nilai *average with distance* yang mengalami penurunan secara signifikan dan berbentuk siku dapat digunakan untuk menentukan nilai *cluster* terbaik.

Berikut merupakan hasil pengujian metode Elbow pada 34 Data. Pada metode ini uji k dimulai dari k=1 hingga k=10. Berikut pada gambar 9. Merupakan hasil penghitungan nilai *Average Within Distance* pada 34 data. Terdapat *cluster* optimal yaitu 2 karena memiliki nilai *average within distenace* tertinggi sebesar 0,202

k	Averag...
1	0.401
2	0.202
3	0.083
4	0.030
5	0.020
6	0.015
7	0.011
8	0.006
9	0.004
10	0.003

Gambar 8. Hasil Perhitungan Average Within Distance

F. Deployment

Pada bagian ini disampaikan hasil laporan berupa klasterisasi dalam menentukan jumlah tingkat Pendidikan yang tingi dan rendah pada setiap Provinsi di Indonesia.

TABEL VI  
Data Hasil Cluster

Cluster 0	Jawa Timur, Jawa Barat, Jawa Tengah
Cluster 1	Sumatera Utara, Sulawesi Selatan, Banten, Nusa Tenggara Timur, Lampung, Sumatera Selatan, Sumatera Barat, Aceh, Riau, Nusa Tenggara Barat, Kalimantan Barat, D.KI. Jakarta, Kalimantan Selatan, D.I Yogyakarta, Sulawesi Tengah, Jambi, Kalimantan Tengah, Sulawesi Tenggara, Sulawesi Utara, Kalimantan Timur, Papua, Bali, Maluku, Bengkulu, Sulawesi Barat, Maluku Utara, Kepulauan Riau, Gorontalo, Papua Barat, Kepulauan Bangka Belitung, Kalimantan Utara, Jawa Timur, Jawa Barat, Jawa Tengah

V. PENUTUP

Berdasarkan rumusan masalah dan tujuan penelitian dapat diambil kesimpulan bahwa klasterisasi Provinsi di Indonesia Berdasarkan Data Pokok Pendidikan Nasional dapat menghasilkan berupa 2 *cluster* optimal yaitu *cluster* 0 Tingkat Pendidikan Tinggi dengan jumlah 3 Provinsi, dan *cluster* 1 Tingkat Pendidikan Tinggi dengan jumlah 31 Provinsi. Dengan nilai Davies Bouldin (DB) sebesar 0.401.



REFERENSI

- [1] R. Kurniawan, M. M. M. Mukarrobil, and M. Mahradianur, "Klasterisasi Tingkat Pendidikan Di Dki Jakarta Pada Tingkat Kecamatan Menggunakan Algoritma K-Means," *Technol. J. Ilm.*, vol. 12, no. 4, p. 234, 2021, doi: 10.31602/tji.v12i4.5633.
- [2] A. B. Subandoro, "Perkembangan Pendidikan di Indonesia," *Kompas*, 2019.  
<https://www.kompasiana.com/aswinbimos13/54f983bfa33311f1068b52ba/perkembangan-pendidikan-indonesia>
- [3] P. I. Sari, "Peran Pendidik dalam Implementasi Media Pembelajaran Terhadap Peserta Didik Generasi 4.0," *Pros. Semin. Nas. Pendidik. FKIP*, vol. 2, no. 1, pp. 508–517, 2019.
- [4] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 2, pp. 337–350, Mar. 2022, doi: 10.30812/matrik.v21i2.1411.
- [5] E. Y. T. P. Dewi and I. Kamila, "Pengelompokan Wilayah Berdasarkan Faktor Pendukung Pendidikan Dengan Jumlah Sekolah Dan Jumlah Guru Menggunakan Algoritma K-Means," *Interval J. Ilm. Mat.*, vol. 2, no. 1, pp. 1–12, 2022, doi: 10.33751/interval.v2i1.5161.
- [6] M. A. Amri, A. P. Windarto, A. Wanto, and I. S. Damanik, "Analisis Metode K-Means Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas Yang Dimiliki Desa," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 674–679, 2019, doi: 10.30865/komik.v3i1.1677.
- [7] M. R. Sholihin and Rudiman, "Pemetaan Sekolah Muhammadiyah di Kabupaten PPU Berdasarkan Fasilitas, Pendidik dan Tenaga Pendidik Menggunakan Metode K-Means Clustering," *J. Keilmuan dan Apl. Tek. Inform.*, vol. 5, no. 36, pp. 1–7, 2022, [Online]. Available: <https://jurnal.yudharta.ac.id/v2/index.php/EXPLORE-IT/article/view/2938/2125>
- [8] D. A. Alodia, A. P. Fialine, D. Endriani, and E. Widodo, "Implementasi Metode K-Medoids Clustering untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan," *Sepren*, vol. 2, no. 2, pp. 1–13, 2021, doi: 10.36655/sepren.v2i2.606.
- [9] P. Meilina, "Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi," *J. Teknol. Univ. Muhammadiyah Jakarta*, vol. 7, no. 1, pp. 11–20, 2015, [Online]. Available: [jurnal.ftumj.ac.id/index.php/jurtek](http://jurnal.ftumj.ac.id/index.php/jurtek)
- [10] Y. A. Wijaya *et al.*, "K-MEANS Di Sekolah Menengah Kejuruan Wahidin Kota Cirebon," vol. 6, no. 2, pp. 552–559, 2022.
- [11] Y. Suhandi, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.
- [12] I. Kamila, U. Khairunnisa, and M. Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, p. 119, 2019, doi: 10.24014/rmsi.v5i1.7381.
- [13] A. A. Hussein, "Improve The Performance of K-means by using Genetic Algorithm for Classification Heart Attack," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 2, p. 1256, 2018, doi: 10.11591/ijece.v8i2.pp1256-1261.
- [14] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain," vol. 6, no. 1, pp. 48–53, 2020.