ANALISA DATA MINING UNTUK KLASTERISASI VAKSIN COVID-19 MENGGUNAKAN ALGORITMA *K-MEANS*

Afif Fathurohman[1], Ahmad Turmudi Zy[2], Edora[3]

Program Studi Teknik Informatika Fakultas Teknik Universitas Pelita Bangsa
Afiffathurohman.rj@gmail.com, afiffathurohman84@gmail.com

Abstract

Corona Virus Disease 2019 (COVID-19) *is a new disease that was reported in Wuhan China in December 2019. To analyze vaccine administration data by region. The k-means algorithm is an easy and effective algorithm to find clusters in the data. The stages in the clustering process and finding the cluster centroid and updating the location of each cluster into a new centroid value in each k cluster repeats steps 3 to 5 until convergence or stops. Analysis of the need for vaccination based on priority area levels by utilizing data on the distribution of* COVID-19 *cases in each province in Indonesia can produce a cluster model using the clustering method according to the specified stages. The K-Means algorithm is used to demonstrate a new knowledge and perspective, namely by grouping provinces in Indonesia based on the 3 proposed clusters, namely cluster 1 is a high priority level consisting of 2 provinces, cluster 2 is the main priority consisting of 2 provinces, and the third is cluster 3 which is a general priority consisting of 30 provinces in Indonesia. Testing using the RapidMiner Studio application also produces relevant insights into the model specified during manual calculations which has a Davies-Bouldin Index evaluation value for the algorithm model of* 0.083.

Keywords: Data Mining, K-Means, Covid-19, *All Indonesian Province*

Abstrak

Corona Virus Disease 2019 (COVID-19) merupakan penyakit baru yang dilaporkan di Wuhan China pada Desember 2019. Untuk menganalisis data pemberian vaksin berdasarkan wilayah. Algoritma k-means merupakan algoritma yang mudah dan efektif untuk menemukan klaster dalam data. Adapun tahapan dalam proses clustering dan menemukan centroid klaster dan memperbarui lokasi masing-masing cluster ke dalam nilai centroid yang baru pada masing-masing k cluster mengulang langka 3 sampai 5 sampai konvergensi atau terhenti. Analisis kebutuhan penyelenggaraan vaksinasi berdasar tingkat wilayah prioritas dengan memanfaatkan data sebaran kasus COVID-19 di setiap provinsi di Indonesia dapat menghasilkan suatu model klaster dengan menggunakan metode klasterisasi sesuai dengan tahapan yang ditentukan. Algoritma K-Means digunakan untuk menunjukkan sebuah pengetahuan dan perspektif baru, yakni dengan mengelompokkan provinsi di Indonesia berdasarkan 3 klaster yang diusulkan, yakni klaster 1 merupakan tingkat prioritas tinggi yang terdiri dari 2 provinsi, klaster 2 adalah prioritas utama yang terdiri dari 2 provinsi, dan yang ketiga adalah klaster 3 yakni prioritas umum yang terdiri dari 30 provinsi di Indonesia. Pengujian menggunakan aplikasi RapidMiner Studio juga menghasilkan wawasan yang relevan terhadap model yang ditentukan saat perhitungan manual yang dimana memiliki nilai evaluasi *Davies-Bouldin Index* untuk model algoritma tersebut sebesar 0,083.

Kata kunci: Data Mining, K-Means, Covid-19, Seluruh Provinis Indonesia

1. Pendahuluan

Corona Virus Disease 2019 (COVID-19) adalah penyakit baru yang dilaporkan di Wuhan China pada Desember 2019 [1]. Meskipun pusat penyebaran virus tersebut berada di Kota Wuhan, China, kini virus tersebut telah tersebar menjangkit keseluruh masyarakat dunia, berdasarkan sumber dari www.covid19.go.id di Indonesia sendiri per tanggal 20 Januari 2021 terdapat 939 ribu lebih kasus dan sebanyak lebih dari 26 ribu lebih kasus yang mengakibatakan kematian. Dilihat dari pesatnya penyebaran COVID- 19 serta bahaya yang ditimbulkan bila tidak lekas ditangani, pemerintah Indonesia mengembangkan vaksin sebagai salah satu cara menghambat penyebaran. Vaksin tidak hanya melindungi mereka yang telah divaksinasi namun juga akan melindungi masyarakat luas dengan mengurangi penyebaran penyakit yang terjadi didalam populasi tersebut [2]. Pengembangan vaksin yang aman dan efektif sangat penting dilakukan karena diharapkan dapat menghentikan penyebaran dan mencegah penyebaran penyakit di masa mendatang.

Dalam menyikapi hal tersebut, Pemerintah Indonesia juga turut aktif dalam rencana kegiatan vaksinasi yang akan diberikan kepada masyarakatnya. Rencana kegiatan vaksinasi

tersebut haruslah mempertimbangkan segala aspek, mulai dari aspek kelayakan vaksin yang akan digunakan, resiko pasca pemakaian, sampai tahapan & prosedur dari pemberian vaksin hingga nantinya sampai ke masyarakat. Semua aspek tersebut haruslah dipertimbangkan secara terperinci agar rencana kegiatan vaksinasi dapat berjalan dengan baik dan terhindar dari hal-hal yang justru akan merugikan [3].

Salah satu aspek yang harus diperhatikan dalam pemberian vaksin covid-19 kepada masyarakat adalah penggunaan vaksin berdasarkan wilayah. Oleh karena itu data pemberian vaksin harus diklasterisasi dengan baik. Salah satu algoritma yang dapat digunakan untuk klasterisasi data yaitu algoritma *k-means*. Klasterisasi atau yang lebih dikenal dengan analisis klaster merupakan proses pengelompokan data ke dalam satu kelas objek yang sama. Sedangkan *k-means*, merupakan metode penganalisaan data dalam *data mining* dimana proses pemodelan tanpa supervisi dan merupakan salah satu metode yang mengelompokan data secara partisi. Pada metode *k-means*, data dikelompokan menjadi beberapa kelompok dimana setiap kelompok mempunyai karakteristik yang mirip atau sama dengan lainnya, namun dengan kelompok lainnya memiliki karakteristik yang berbeda [4].

Pada penelitian ini akan dilakukan klasterisasi data covid-19 menggunakan algoritma *k-means* untuk mengetahui vaksin yang telah diberikan maupun yang belum diberikan kepada masyarakat berdasarkan wilayah untuk mempermudah dalam pemerataan program vaksinasi yang diselenggarakan oleh pemerintah.

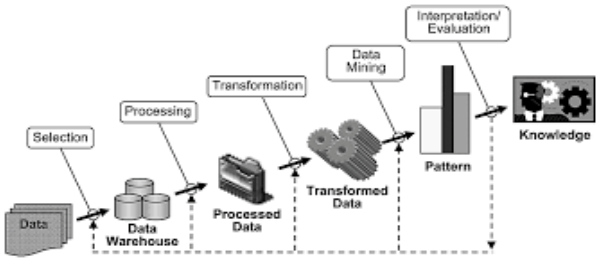
2. Metode Penelitian

2.1 Objek Penelitian

Pada penelitian ini, tahapan yang akan digunakan dalam melakukan *cluster* data vaksin terhadap vaksinisasi masyarakat dan penentuan atribut standar penilian dari data vaksin untuk mempermudah penelitian sehingga penelitian dapat berjalan dengan baik dan sistematis, serta memenuhi tujuan yang diinginkan.

2.2 Tahapan Penelitian

Dalam melakukan analisa dan mencari pola data untuk dijadikan sebuah dataset dalam memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan maka dibuat langkah – langkah dalam tahapan penelitian yang akan dilakukan berikut :



Gambar 3.1 Tahap Penelitian [10]

2.2.1 Pengumpulan Data

Pengumpulan data cara yang dapat digunakan oleh peneliti untuk mengumpulkan data kasus Covid-19 di Indonesia yang memiliki atribut provinsi, kasus terkonfirmasi dan kasus sembuh yang ditetapkan dalam penelitian sehingga dapat dipelajari kemudian ditarik kesimpulannya. Dataset yang digunakan dalam penelitian diambil dari *site* https://www.antaranews.com/covid-19 dengan status data per **tanggal 25 Maret 2022**.

2.2.2 Pengolahan Data

Pada tahap ini menjelaskan tentang tahap awal data mining. Data yang telah didapatkan akan diolah keformat yang dibutuhkan, pengelompokkan dan penentuan atribut data. Dalam melakukan pengolahan data awal, akan dilakukan beberapa tahapan agar didapatkan data yang bisa digunakan untuk tahap selanjutnya. Beberapa tahapan tersebut yaitu :

2.2.2.1 *Data Cleaning*

Pada tahap data cleaning merupakan proses pembersihan dari data yang akan dipakai untuk penghapusan data dengan membuang *missing value*, duplikasi data, dan memeriksa inkonsistensi dari sebuah data dan memperbaiki kesalahan pada data. Proses pemberihan data dilakukan secara manual untuk memastikan bahwa data yang telah dipilih layak untuk dilakuka proses permodelan.

Tabel 3.1 Contoh *Data Cleaning*

Provinsi	Dirawat	Terkonfirmasi	Sembuh	Meninggal
SUMATERA_SELATAN	5	11	6	0
BENGKULU	1	1	0	0
PROVINSI JAWA TENGAH	0	1	1	0
LAMPUNG TENGAH	1	1	0	0

2.2.2.2 Data Selection

Tahapan *data selection* merupakan proses pemilihan data dari sekumpulan data operasional yang ada sebelum masuk ke tahap mining data

maupun informasi. Pada tahap ini juga akan dilakukan pemilihan data dengan seleksi atribut yang diperlukan ke proses selanjutnya, sehingga atribut yang tidak memiliki keterkaitan terhadap tahapan penerapan model nantinya akan direduksi agar dimensi data menjadi lebih ringkas.
Pada dataset awal terdapat 5 atribut seperti yang ada pada Tabel 3.2 berikut.

Tabel 3.2 Atribut awal pada dataset

No.	Nama Atribut	Tipe Data
1	Provinsi	Text
2	Dirawat	Numerik
3	Terkonfirmasi	Numerik
4	Sembuh	Numerik
5	Meninggal	Numerik

Pada proses seleksi atribut ini beberapa atribut yang tidak berhubungan dan berkaitan dengan tahap pemodelan klastering dan algoritma K-Means tidak akan digunakan dalam proses selanjutnya. Atribut yang dipakai hanya atribut **provinsi, terkonfirmasi**, dan **meninggal**.

Tabel 3.3 Hasil proses seleksi atribut

No.	Nama Atribut	Tipe Data
1	Provinsi	Text
2	Terkonfirmasi	Numerik
3	Meninggal	Numerik

2.2.2.3 *Data Transformation*

Tahap *data transformation* merupakan proses mengubah format data awal menjadi sebuah format data standar untuk proses pembacaan data dengan algoritma pada program maupun aplikasi yang digunakan.

Tabel 3.4 Contoh data *transformation*

Nama Prov	Kasus Konfirmasi	Kasus Sembuh
Aceh	43.315	38.906
Sumatera Utara	153.990	146.280
Sumatera Barat	103.181	96.188
Riau	150.006	142.874
Jambi	37.876	35.966
Sumatera Selatan	80.077	74.259
Bengkulu	28.978	26.866

2.2.3 Model Data Mining da Algoritma

Pada tahap ini dilakukan pemodelan data, metode yang dipakai pada penelitian ini adalah metode klasterisasi dengan algoritma *k-means* yang merupakan sebuah klasterisasi sederhana yang mengaplikasikan dengan asumsi ke tidak tergantungan (independen) yang tinggi. *K-means* merupakan metode clustering sederhana berdasarkan data dimana dilakukan melalui training set sejumlah data secara efisien dan mengasumsikan bahwa nilai dari sebuah input atribut pada kelas yang diberikan tidak tergantung dengan nilai atribut yang lain.

Algoritma *K-means* merupakan algoritma yang mudah dan efektif untuk menemukan klaster dalam data. Adapun tahapan dalam proses clustering dengan k-means antara lain[11].

1. Menentukan berapa banyak data yang akan dipartisi.
2. Secara acak menetapkan k record menjadi lokasi cluster center awal.
3. Menemukan pusat cluster terdekat. Jadi, dalam arti tertentu, masing-masing cluster center "memiliki" subset dari catatan, sehingga mewakili sebuah partisi dari kumpulan data Oleh karena itu kami memiliki k cluster, C1, C2,...Ck.
4. Menemukan centroid klaster dan memperbarui lokasi masing-masing cluster ke dalam nilai centroid yang baru pada masing-masing k cluster
5. Mengulang langka 3 sampai 5 sampai konvergensi atau terhenti. Kriteria terdekat pada langkah ketiga,biasanya jarang Euclidean, walaupun kriteria lainnya dapat diterapkan. *Cluster* centroid pada langkah 4 ditemukan sebagai berikut. Misalkan terdapat nilai n titik data (a1, b1, c1), (a2, b2, c2) ,..., (an,bn,cn), centroid dari masing-masing titik ini adalah proses gravitasi titik dan terletak di titik (∑a i /n, ∑b i /n, ∑c i /n).

Algoritma k-means akan berakhir apabila centroid tidak lagi berubah. Maksudnya, alhgoritma berakhir ketika semua klaster C1, C2, C3 ….., Ck. Semua caatanyang dimiliki oleh masing-masing cluster centermain di klaster tersebut. Dan juga algoritma k-means akan berakhir ketika beberapa kriteria konvergensi dipenuhi, seperti tidak ada penyusutan signifikan dalam jumlah kesalahan kuadrat[11]

[1]

$$\text{Vij} = \frac{1}{Ni}\sum_{k=0}^{Ni} \text{Xkj}$$

[2]

$$D = \sqrt{(Xi - Si)^2 + (yi - ti)^2}$$

Keterangan :

D = Euclidean Distance

i = banyaknya objek

(x,y) = koordinat objek

(s,t) = koordinat centroid atribut (l)

2.2.4 Evaluasi Pengujian

Evaluasi dapat dilakukan dengan cara mengamati dan menganalisa hasil dari algoritma yang digunakan untuk memastikan bahwa hasil pengujian benar-benar sesuai dengan pembahasan. Melakukan pengecekan terhadap setiap nilai atribut dan model yang sudah dibangun. Kemudian melakukan evaluasi dengan cara mengamati dan menganalisa hasil dari algoritma yang digunakan untuk memastikan bahwa hasil pengujian menghasilkan nilai *cluster* benar dan sesuai hasil pembahasan, pengujian dilakukan untuk mengukur keakuratan data hasil dari tiap model yang diusulkan.

2.3 Kebutuhan Perangkat Lunak dan Perangkat Keras

Analisis kebutuhan perangkat lunak dan keras merupakan yang dibutuhkan untuk menentukan spesifikasi kebutuhan penelitian. Spesifiikasi ini juga meliputi elemen-elemen atau komponen-komponenapa saja yang dibutuhkan untuk penelitian yang akan dibangun sampai dengan tahap evaluasi.

2.3.1 Kebutuhan Perangkat Lunak *(Software)*

Perangkat lunak, versi dan fungsi dapat dilihat pada Tabel 3.5 dibawah ini.

Tabel 3.5 Perangkat Lunak *(Software)*

Software	**Versi**	**Fungsi**
Sistem Operasi *Microsoft Windows*	10	Sebagai sistem operasi penelitian ini
Microsoft Office Word	2016	Digunakan untuk mengolah laporan ini
Microsoft Office Excel	2016	Digunakan sebagai media penulisan, dan pengolahan *dataset*
Rapidminer	38.4	Digunakan untuk mengolah *dataset* dan untuk melihat hasil akurasi dari algoritma yang digunakan

2.3.2 Kebutuhan Perangkat Keras *(Hardware)*

Selain perangkat lunak (*software*) dibutuhkan pula perankat keras (*Hardware*) sebagai pendukung penelitian *data mining* , yaitu laptop. Adapun spesifikasi laptop dijelaskan pada tabel 3.5 dibawah ini :

Tabel 3. 6 Perangkat Keras *(Hardware)*

Spesifikasi Hardware	**Keterangan**
Processor	Intel ® Core (TM) i5-7200U CPU @2.50 GHz
RAM	8,00 GB
System Operation	Windows 10 Enterprise 64-bit (10.0, Build 16299)

3. Hasil dan Pembahasan

3.1 Analisis Dataset Covid-19 Provinsi di Informatika

Analisis data dilakukan pada penelitian ini adalah mengumpulkan serta mempersiapkan data, yang dimana dataset penelitian ini merupakan kumpulan data akumulasi kasus harian Covid-19 di setiap provinsi di Indonesia per tanggal 25 Maret 2022 yang diambil dari situs resmi berita Antara pada *site* berikut **https://www.antaranews.com/covid-19**, dimana data memiliki total baris sebanyak 38 *record* data. Dalam tahapan *pra-processing data*, dilakukan pembersihan data yang inkonsisten sebanyak 4 baris *record data*, dan dilakukan seleksi atribut untuk mereduksi dimensi data dengan hanya menggunakan jenis atribut yang akan dimodelkan, sehingga dataset yang akan diproses pada tahapan selanjutnya hanya memiliki 34 *record data* dan 3 atribut seperti terlihat pada tabel dibawah ini :

Tabel 4. 1 Dataset Covid-19 per Provinsi di Indonesia

No	Nama Prov	Kasus Konfirmasi	Kasus Sembuh
1	Aceh	43.315	38.906
2	Sumatera Utara	153.990	146.280
3	Sumatera Barat	103.181	96.188
4	Riau	150.006	142.874
5	Jambi	37.876	35.966
6	Sumatera Selatan	80.077	74.259
7	Bengkulu	28.978	26.866
8	Kepulauan Bangka Belitung	64.915	61.441
9	Lampung	73.308	57.156
10	Kepulauan Riau	69.337	63.680
11	DKI Jakarta	1.231.444	1.202.905
12	Jawa Barat	1.091.786	1.023.684
13	Jawa Tengah	620.629	565.091
14	Daerah Istimewa Yogyakarta	218.190	184.920
15	Jawa Timur	571.020	535.001
16	Banten	286.512	269.392
17	Bali	155.825	150.213
18	Nusa Tenggara Barat	35.949	34.747
19	Nusa Tenggara Timur	91.684	82.541
20	Kalimantan Barat	63.059	58.569
21	Kalimantan Tengah	55.156	52.263
22	Kalimantan Selatan	83.984	80.879
23	Kalimantan Timur	205.128	195.603
24	Kalimantan Utara	44.979	41.339
25	Sulawesi Utara	51.312	49.232
26	Sulawesi Tengah	59.982	54.829
27	Sulawesi Selatan	142.746	136.407
28	Sulawesi Tenggara	25.455	24.464
29	Gorontalo	13.803	12.863
30	Sulawesi Barat	15.381	14.336
31	Maluku	18.517	18.137
32	Maluku Utara	14.348	13.220
33	Papua	47.985	35.578
34	Papua Barat	31.131	30.325

3.2 Pemodelan dan Contoh Hasil Perhitungan Algoritma K-Means

Pada tahap ini penulis menentukan jumlah *cluster* yang akan dipakai, yaitu akan dibentuk sebanyak 3 *cluster*, antara lain C1, C2, dan C3. Dengan adanya 3 *cluster* ini maka akan setiap kelompok klaster juga akan terbagi menjadi 3 label kelompok, yakni kelompok Prioritas Utama pada C1, Prioritas Tinggi pada C2, dan Prioritas Umum pada C3.

Berikutnya ditentukan titik awal dari setiap *cluster*. Pada penelitian ini untuk titik awal *cluster* diambil secara random dengan menghitung nilai maksimal yang terdapat pada data ke 1 sampai data ke 11 untuk titik awal *cluster* ke 1, kemudian nilai maksimal yang terdapat pada data ke 12 sampai data ke 21 untuk titik awal *cluster* ke 2 dan nilai maksimal yang terdapat pada data ke 22 sampai data ke 34 untuk titik awal *cluster* ke 3 dari masing-masing atribut kasus terkonfirmasi dan kasus sembuh, sehingga hasil penentukan titik awal untuk tahap perhitunga iterasi ke 1 adalah pada tabel 4.2 berikut.

Tabel 4. 2 Titik awal klaster untuk proses iterasi ke-1

Titik Pusat Awal	Kasus Konfirmasi	Kasus Sembuh
Cluster ke-1	1.231.444	1.202.905
Cluster ke-2	1.091.786	1.023.684
Cluster ke-3	205.128	195.603

Dalam menghitung jarak terdekat dari masing – masing data terhadap nilai awal *cluster* ialah dengan menggunakan rumus *Euclidean Distance* seperti dalam persamaan berikut :

$$D = \sqrt{(xi - si)^2 + (yi - ti)^2} \quad [1]$$

Keterangan :

D = *Euclidean Distance*

i = banyaknya objek

(x,y) = koordinat objek

(s,t) = koordinat centroid

Berikut merupakan contoh dari perhitungan iterasi pertama pada baris data ke-1 , ke-2, dan ke-3 pada dataset yang ada.

1. *Euclidean Distance* pada Data Ke- 1

D1(c1) =

$\sqrt{(43.315-1.233.144)^2+(38.906-1.202.905)^2} = 1.663.293$

D1(c2) =

$\sqrt{(43.315-1.091.786)^2+(38.906-1.203.684)^2} = 1.438.429$

D1(c3) =

$\sqrt{(43.315-205.128)^2+(38.906-195.603)^2} = 225.250$

Jarak terpendek pada data pertama adalah 225.250 yang ada pada klaster ke-3 (C3).

2. *Euclidean Distance* pada Data Ke- 2

D2(c1) =

$\sqrt{(153.990-1.233.144)^2+(146.280-1.202.905)^2} = 1.509.094$

D2(c2) =

$\sqrt{(153.990-1.091.786)^2+(146.280-1.203.684)^2} = 1.284.250$

D2(c3) =

$\sqrt{(153.990-205.128)^2+(146.280-195.603)^2} = 71.048$

Jarak terpendek pada data kedua adalah 71.048 yang ada pada klaster ke-3 (C3).

3. *Euclidean Distance* pada Data Ke- 3

D3(c1) =

$\sqrt{(103.181-1.233.144)^2+(96.188-1.202.905)^2} = 1.580.443$

D3(c2) =

$\sqrt{(103.181-1.091.786)^2+(96.188-1.203.684)^2} = 1.355.577$

D3(c3) =

$\sqrt{(103.881-205.128)^2+(96.188-195.603)^2} = 142.396$

Jarak terpendek pada data ketiga adalah 142.396 yang ada pada klaster ke-3 (C3).

Pada proses iterasi pertama tersebut dapat dilihat bahwa jarak terdekat yang dimiliki oleh data ke-1, data ke-2 dan data ke-3 ada di kelompok klaster ke-3 yakni (C3). Pada tabel berikut dapat dilihat untuk hasil dari perhitungan *Euclidean Distance* pada tahap iterasi pertama.

Tabel 4. 3 Hasil perhitungan *Euclidean Distance* pada proses iterasi ke-1

Nama Prov	C1	C2	C3	Jarak Terpendek
Aceh	1.663.293	1.438.429	225.250	225.250
Sumatera Utara	1.509.094	1.284.250	71.048	71.048
Sumatera Barat	1.580.443	1.355.577	142.396	142.396
Riau	1.514.323	1.289.487	76.281	76.281
Jambi	1.669.237	1.444.408	231.208	231.208
Sumatera Selatan	1.612.293	1.387.430	174.247	174.247
Bengkulu	1.681.961	1.457.123	243.928	243.928
Kepulauan Bangka Belitung	1.632.094	1.407.258	194.060	194.060
Lampung	1.629.116	1.404.092	191.165	191.165
Kepulauan Riau	1.627.368	1.402.501	189.322	189.322
DKI Jakarta	-	227.210	1.438.048	-
Jawa Barat	227.210	-	1.213.211	-
Jawa Tengah	883.120	657.493	556.024	556.024
Daerah Istimewa Yogyakarta	1.436.307	1.211.072	16.874	16.874
Jawa Timur	939.285	714.149	499.067	499.067
Banten	1.328.286	1.103.369	109.855	109.855
Bali	1.505.030	1.280.224	67.015	67.015
Nusa Tenggara Barat	1.671.467	1.446.647	233.444	233.444
Nusa Tenggara Timur	1.598.208	1.373.300	160.164	160.164
Kalimantan Barat	1.635.429	1.410.577	197.388	197.388
Kalimantan Tengah	1.645.488	1.420.655	207.456	207.456
Kalimantan Selatan	1.604.870	1.380.053	166.846	166.846
Kalimantan Timur	1.438.048	1.213.211	-	-
Kalimantan Utara	1.660.402	1.435.551	222.363	222.363
Sulawesi Utara	1.650.355	1.425.532	212.330	212.330
Sulawesi Tengah	1.640.244	1.415.380	202.200	202.200
Sulawesi Selatan	1.524.035	1.299.206	85.998	85.998
Sulawesi Tenggara	1.686.159	1.461.336	248.135	248.135
Gorontalo	1.702.601	1.477.771	264.574	264.574
Sulawesi Barat	1.700.443	1.475.612	262.415	262.415
Maluku	1.695.543	1.470.725	257.523	257.523
Maluku Utara	1.701.962	1.477.129	263.933	263.933
Papua	1.662.296	1.437.315	224.281	224.281
Papua Barat	1.678.003	1.453.187	239.983	239.983

Berikutnya jarak dari hasil perhitungan tersebut akan dilakukan perbandingan diantara 3 *cluster* data terdekat dengan titik pusat *cluster* dengan mengambil nilai terkecil, jarak ini menunjukkan bahwa data tersebut berada dalam 1 kelompok dengan pusat *cluster* terdekat. Berikut adalah matriks kelompok *cluster* yang terjadi pada iterasi ke-1 yang dilambangkan dengan notasi huruf "T" dilihat pada tabel 4.4.

Tabel 4.4 Matriks kelompok data pada proses iterasi ke-1

Nama Prov	C1	C2	C3
Aceh			T
Sumatera Utara			T
Sumatera Barat			T
Riau			T
Jambi			T
Sumatera Selatan			T
Bengkulu			T
Kepulauan Bangka Belitung			T
Lampung			T
Kepulauan Riau			T
DKI Jakarta	T		
Jawa Barat		T	
Jawa Tengah			T
Daerah Istimewa Yogyakarta			T
Jawa Timur			T
Banten			T
Bali			T
Nusa Tenggara Barat			T
Nusa Tenggara Timur			T
Kalimantan Barat			T
Kalimantan Tengah			T
Kalimantan Selatan			T
Kalimantan Timur			T
Kalimantan Utara			T
Sulawesi Utara			T
Sulawesi Tengah			T
Sulawesi Selatan			T
Sulawesi Tenggara			T
Gorontalo			T
Sulawesi Barat			T
Maluku			T
Maluku Utara			T
Papua			T
Papua Barat			T
Index Total (T)	**1**	**1**	**32**

Penentuan kembali titik *cluster* baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok *cluster* menggunakan rumus berikut ini :

[2]

$$Vij = \frac{1}{Ni}\sum_{k=0}^{Ni} Xkj$$

Keterangan :

*V*ij = *Centroid* rata-rata *cluster* ke-*i* untuk variabel ke-*i*

*N*i = jumlah *cluster* ke -i

i, *k* = indeks dari cluster

j = indeks dari variabel

Xkj = nilai data ke –k variabel ke –j dalam cluster

Berikut adalah perhitungan untuk penentuan nilai titik *cluster* baru untuk proses iterasi kedua :

C1 baru ke 2 (Kasus Konfirmasi)

$$= \frac{1.231.444}{1} = 1.231.444$$

C1 baru ke 2 (Kasus Sembuh)

$$= \frac{1.202.905}{1} = 1.202.905$$

C2 baru ke 2 (Kasus Konfirmasi)

$$= \frac{1.091.786}{1} = 1.091.786$$

C2 baru ke 2 (Kasus Sembuh)

$$= \frac{1.023.684}{1} = 1.023.684$$

C3 baru ke 2 (Kasus Konfirmasi)

$$= \frac{3.657.758}{32} = 114.305$$

C3 baru ke 2 (Kasus Sembuh)

$$= \frac{3.383.565}{32} = 105.736$$

Tabel 4. 5 Titik awal klaster untuk proses iterasi ke-2

Titik Pusat Awal	Kasus Konfirmasi	Kasus Sembuh
Cluster ke-1	1.231.444	1.202.905
Cluster ke-2	1.091.786	1.023.684
Cluster ke-3	114.305	105.736

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-2.

Berikut merupakan contoh dari perhitungan iterasi kedua pada baris data ke-1 , ke-2, dan ke-3 pada dataset yang ada.

1. *Euclidean Distance* pada Data Ke- 1

D1(c1) =

$$\sqrt{(43.315-1.233.144)^2+(38.906-1.202.905)^2} = 1.663.293$$

D1(c2) =

$$\sqrt{(43.315-1.091.786)^2+(38.906-1.203.684)^2} = 1.438.429$$

D1(c3) =

$$\sqrt{(43.315-114.305)^2+(38.906-105.736)^2} = 97.498$$

Jarak terpendek pada data pertama adalah 97.498 yang ada pada klaster ke-3 (C3).

2. a *Euclidean Distance* pada Data Ke- 2

D2(c1) =

$$\sqrt{(153.990-1.233.144)^2+(146.280-1.202.905)^2} = 1.509.094$$

D2(c2) =

$$\sqrt{(153.990-1.091.786)^2+(146.280-1.203.684)^2} = 1.284.250$$

D2(c3) =

$$\sqrt{(153.990-114.305)^2+(146.280-105.736)^2} = 56.733$$

Jarak terpendek pada data kedua adalah 56.733 yang ada pada klaster ke-3 (C3).

3. a *Euclidean Distance* pada Data Ke- 3

D3(c1) =

$$\sqrt{(103.181-1.233.144)^2+(96.188-1.202.905)^2} = 1.580.443$$

D3(c2) =

$$\sqrt{(103.181-1.091.786)^2+(96.188-1.203.684)^2} = 1.355.577$$

D3(c3) =

$$\sqrt{(103.881-114.305)^2+(96.188-105.736)^2} = 14.660$$

Jarak terpendek pada data ketiga adalah 14.660 yang ada pada klaster ke-3 (C3).

Pada proses iterasi kedua tersebut dapat dilihat bahwa jarak terdekat yang dimiliki oleh data ke-1, data ke-2 dan data ke-3 ada di kelompok klaster ke-3 yakni (C3). Pada tabel berikut dapat dilihat untuk hasil dari perhitungan *Euclidean Distance* pada tahap iterasi kedua.

Tabel 4. 6 Hasil perhitungan *Euclidean Distance* pada proses iterasi ke-2

Nama Prov	C1	C2	C3	Jarak Terpendek
Aceh	1.663.293	1.438.429	97.498	97.498
Sumatera Utara	1.509.094	1.284.250	56.733	56.733
Sumatera Barat	1.580.443	1.355.577	14.660	14.660
Riau	1.514.323	1.289.487	51.515	51.515
Jambi	1.669.237	1.444.408	103.486	103.486
Sumatera Selatan	1.612.293	1.387.430	46.501	46.501
Bengkulu	1.681.961	1.457.123	116.195	116.195
Kepulauan Bangka Belitung	1.632.094	1.407.258	66.343	66.343
Lampung	1.629.116	1.404.092	63.567	63.567
Kepulauan Riau	1.627.368	1.402.501	61.570	61.570
DKI Jakarta	-	227.210	1.565.816	-
Jawa Barat	227.210	-	1.340.931	-
Jawa Tengah	883.120	657.493	683.645	657.493
Daerah Istimewa Yogyakarta	1.436.307	1.211.072	130.622	130.622
Jawa Timur	939.285	714.149	626.783	626.783
Banten	1.328.286	1.103.369	237.568	237.568
Bali	1.505.030	1.280.224	60.845	60.845
Nusa Tenggara Barat	1.671.467	1.446.647	105.731	105.731
Nusa Tenggara Timur	1.598.208	1.373.300	32.400	32.400
Kalimantan Barat	1.635.429	1.410.577	69.648	69.648
Kalimantan Tengah	1.645.488	1.420.655	79.737	79.737
Kalimantan Selatan	1.604.870	1.380.053	39.208	39.208
Kalimantan Timur	1.438.048	1.213.211	127.769	127.769
Kalimantan Utara	1.660.402	1.435.551	94.621	94.621
Sulawesi Utara	1.650.355	1.425.532	84.622	84.622
Sulawesi Tengah	1.640.244	1.415.380	74.448	74.448
Sulawesi Selatan	1.524.035	1.299.206	41.828	41.828
Sulawesi Tenggara	1.686.159	1.461.336	120.414	120.414
Gorontalo	1.702.601	1.477.771	136.843	136.843
Sulawesi Barat	1.700.443	1.475.612	134.685	134.685
Maluku	1.695.543	1.470.725	129.804	129.804
Maluku Utara	1.701.962	1.477.129	136.201	136.201
Papua	1.662.296	1.437.315	96.543	96.543
Papua Barat	1.678.003	1.453.187	112.271	112.271

Berikutnya jarak dari hasil perhitungan tersebut akan dilakukan perbandingan diantara 3 *cluster* data terdekat dengan titik pusat *cluster* dengan mengambil nilai terkecil, jarak ini menunjukkan bahwa data tersebut berada dalam 1 kelompok dengan pusat *cluster* terdekat. Berikut adalah matriks kelompok *cluster* yang terjadi pada iterasi ke-3 yang dilambangkan dengan notasi huruf "T" dilihat pada tabel 4.10.

Tabel 4. 10 Matriks kelompok data pada proses iterasi ke-3

Nama Prov	C1	C2	C3
Aceh			T
Sumatera Utara			T
Sumatera Barat			T
Riau			T
Jambi			T
Sumatera Selatan			T

Bengkulu			T
Kepulauan Bangka Belitung			T
Lampung			T
Kepulauan Riau			T
DKI Jakarta	T		
Jawa Barat	T		
Jawa Tengah		T	
Daerah Istimewa Yogyakarta			T
Jawa Timur		T	
Banten			T
Bali			T
Nusa Tenggara Barat			T
Nusa Tenggara Timur			T
Kalimantan Barat			T
Kalimantan Tengah			T
Kalimantan Selatan			T
Kalimantan Timur			T
Kalimantan Utara			T
Sulawesi Utara			T
Sulawesi Tengah			T
Sulawesi Selatan			T
Sulawesi Tenggara			T
Gorontalo			T
Sulawesi Barat			T
Maluku			T
Maluku Utara			T
Papua			T
Papua Barat			T
Index Total (T)	**2**	**2**	**30**

Penentuan kembali titik *cluster* baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok *cluster*.

Berikut adalah perhitungan untuk penentuan nilai titik *cluster* baru untuk proses iterasi keempat :

C1 baru ke 2 (Kasus Konfirmasi)

$$= \frac{2.323.230}{2} = 1.161.615$$

C1 baru ke 2 (Kasus Sembuh)

$$= \frac{2.226.589}{2} = 1.113.295$$

C2 baru ke 2 (Kasus Konfirmasi)

$$= \frac{1.191.649}{2} = 595.825$$

C2 baru ke 2 (Kasus Sembuh)

$$= \frac{1.100.092}{2} = 550.046$$

C3 baru ke 2 (Kasus Konfirmasi)

$$= \frac{2.466.109}{30} = 82.204$$

C3 baru ke 2 (Kasus Sembuh)

$$= \frac{2.283.473}{30} = 76.116$$

Tabel 4. 11 Titik awal klaster untuk proses iterasi ke-4

Titik Pusat Awal	Kasus Konfirmasi	Kasus Sembuh
Cluster ke-1	1.161.615	1.113.295
Cluster ke-2	595.825	550.046
Cluster ke-3	82.204	76.116

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-4.

Berikut merupakan contoh dari perhitungan iterasi keempat pada baris data ke-1, ke-2, dan ke-3 pada dataset yang ada.

1. *Euclidean Distance* pada Data Ke- 1

D1(c1) =

$$\sqrt{(43.315-1.161.615)^2+(38.906-1.113.295)^2} = 1.550.776$$

D1(c2) =

$$\sqrt{(43.315-595.285)^2+(38.906-550.046)^2} = 752.682$$

D1(c3) =

$$\sqrt{(43.315-82.204)^2+(38.906-76.116)^2} = 53.823$$

Jarak terpendek pada data pertama adalah 53.823 yang ada pada klaster ke-3 (C3).

2. a *Euclidean Distance* pada Data Ke- 2

D2(c1) =

$$\sqrt{(153.990-1.161.615)^2+(146.280-1.113.295)^2} = 1.396.576$$

D2(c2) =

$$\sqrt{(153.990-595.285)^2+(146.280-550.046)^2} = 598.335$$

D2(c3) =

$$\sqrt{(153.990-82.204)^2+(146.280-76.116)^2} = 100.381$$

Jarak terpendek pada data kedua adalah 100.381 yang ada pada klaster ke-3 (C3).

3. *Euclidean Distance* pada Data Ke- 3

D3(c1) =

$$\sqrt{(103.181-1.161.615)^2+(96.188-1.113.295)^2} = 1.467.920$$

D3(c2) =

$$\sqrt{(103.181-595.285)^2+(96.188-550.046)^2} = 669.839$$

D3(c3) =

$$\sqrt{(103.881-82.204)^2+(96.188-76.116)^2} = 29.034$$

Jarak terpendek pada data ketiga adalah 29.034 yang ada pada klaster ke-3 (C3).

Pada proses iterasi keempat tersebut dapat dilihat bahwa jarak terdekat yang dimiliki oleh data ke-1, data ke-2 dan data ke-3 ada di kelompok klaster ke-3 yakni (C3). Pada tabel berikut dapat dilihat untuk hasil dari perhitungan *Euclidean Distance* pada tahap iterasi keempat.

Tabel 4.12 Hasil perhitungan *Euclidean Distance* pada proses iterasi ke-4

Nama Prov	C1	C2	C3	Jarak Terpendek
Aceh	1.550.776	752.682	53.823	53.823
Sumatera Utara	1.396.576	598.535	100.381	100.381
Sumatera Barat	1.467.920	669.839	29.034	29.034
Riau	1.401.809	603.774	95.152	95.152
Jambi	1.556.736	758.673	59.808	59.808
Sumatera Selatan	1.499.773	701.690	2.823	2.823
Bengkulu	1.569.456	771.383	72.516	72.516
Kepulauan Bangka Belitung	1.519.588	721.526	22.677	22.677
Lampung	1.516.523	718.306	20.943	20.943
Kepulauan Riau	1.514.847	716.757	17.894	17.894
DKI Jakarta	113.605	911.173	1.609.474	113.605
Jawa Barat	113.605	685.792	1.384.609	113.605
Jawa Tengah	770.190	29.011	727.323	29.011
Daerah Istimewa Yogyakarta	1.323.605	525.285	174.157	174.157
Jawa Timur	826.575	29.011	670.460	29.011
Banten	1.215.721	417.661	281.243	281.243
Bali	1.392.530	594.530	104.453	104.453
Nusa Tenggara Barat	1.558.970	760.916	62.055	62.055
Nusa Tenggara Timur	1.485.667	687.545	11.453	11.453
Kalimantan Barat	1.522.915	724.837	25.969	25.969
Kalimantan Tengah	1.532.984	734.922	36.063	36.063
Kalimantan Selatan	1.492.371	694.333	5.085	5.085
Kalimantan Timur	1.325.528	527.516	171.428	171.428
Kalimantan Utara	1.547.890	749.809	50.942	50.942
Sulawesi Utara	1.537.856	739.803	40.952	40.952
Sulawesi Tengah	1.527.725	729.635	30.772	30.772
Sulawesi Selatan	1.411.525	613.496	85.442	85.442
Sulawesi Tenggara	1.573.662	775.602	76.735	76.735
Gorontalo	1.590.101	792.032	93.164	93.164
Sulawesi Barat	1.587.943	789.873	91.006	91.006
Maluku	1.583.049	784.991	86.125	86.125
Maluku Utara	1.589.461	791.389	92.522	92.522
Papua	1.549.724	751.535	53.049	53.049
Papua Barat	1.565.509	767.456	68.595	68.595

Berikutnya jarak dari hasil perhitungan tersebut akan dilakukan perbandingan diantara 3 *cluster* data terdekat dengan titik pusat *cluster* dengan mengambil nilai terkecil, jarak ini menunjukkan bahwa data tersebut berada dalam 1 kelompok dengan pusat *cluster* terdekat. Berikut adalah matriks kelompok *cluster* yang terjadi pada iterasi ke-4 yang dilambangkan dengan notasi huruf "T" dilihat pada tabel 4.13.

Tabel 4. 13 Matriks kelompok data pada proses iterasi ke-4

Nama Prov	C1	C2	C3
Aceh			T
Sumatera Utara			T
Sumatera Barat			T
Riau			T
Jambi			T
Sumatera Selatan			T
Bengkulu			T
Kepulauan Bangka Belitung			T
Lampung			T
Kepulauan Riau			T
DKI Jakarta	T		
Jawa Barat	T		
Jawa Tengah		T	
Daerah Istimewa Yogyakarta			T
Jawa Timur		T	
Banten			T
Bali			T
Nusa Tenggara Barat			T
Nusa Tenggara Timur			T
Kalimantan Barat			T
Kalimantan Tengah			T
Kalimantan Selatan			T
Kalimantan Timur			T
Kalimantan Utara			T
Sulawesi Utara			T
Sulawesi Tengah			T
Sulawesi Selatan			T
Sulawesi Tenggara			T
Gorontalo			T
Sulawesi Barat			T
Maluku			T
Maluku Utara			T

Papua			T
Papua Barat			T
Index Total (T)	**2**	**2**	**30**

Pada perhitungan diatas maka untuk memenuhi syarat indeks kelompok masing – masing klaster memliki kesamaan jumlah objek seperti pada iterasi sebelumnya, maka dengan demikian untuk proses perhitungan iterasi berhenti di iterasi ke-4. Perbandingan indeks jumlah objek data pada masing – masing kelompok klaster dapat dilihat pada tabel 4.14 berikut.

Tabel 4.14 Perbandingan jumlah objek data dari proses iterasi ke-3 & ke-4

Proses Iterasi	Jml_Objek C1	Jml_Objek C2	Jml_Objek C3
Iterasi ke 3	2	2	30
Iterasi ke 4	2	2	30

3.3 Pembentukan Kelompok Klaster

Hasil penerapan metode *cluster*isasi dengan algoritma *K-Means* pada dataset kasus Covid-19 provinsi di Indonesia yang digunakan setelah dimodelkan dengan melalui tahapan yang ditentukan mendapatkan posisi *cluster* optimal pada iterasi ke-4, karena jumlah objek dari masing – masing *cluster* sudah tidak berubah. Berikut merupakan *cluster* beserta anggota objek yang terbentuk pada proses *data mining*:

1. *Cluster* pertama (C1) memiliki pusat *cluster* (1.161.615 & 1.113.295)

   Pada *cluster* ini adalah kelompok objek yang terbentuk dari data kasus Covid-19 dengan kategori propinsi dengan Prioritas Utama. Terdapat 2 *record* data provinsi yang masuk dalam *cluster* pertama, yakni provinsi DKI Jakarta dan provinsi Jawa Barat.

2. *Cluster* kedua (C2) memiliki pusat *cluster* (595.825 & 550.046).

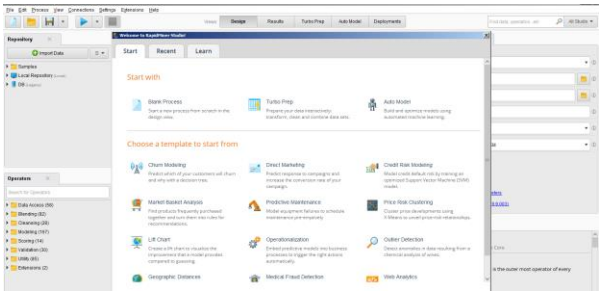
   Pada *cluster* ini adalah kelompok objek yang terbentuk dari data kasus Covid-19 dengan kategori propinsi dengan Prioritas Tinggi. Terdapat 2 *record* data provinsi yang masuk dalam *cluster* kedua, yakni provinsi Jawa Tengah dan provinsi Jawa Timur

3. *Cluster* ketiga (C3) memiliki pusat *cluster* (82.204 & 76.116).

   Pada *cluster* ini adalah kelompok objek yang terbentuk dari data kasus Covid-19 dengan kategori propinsi dengan Prioritas Umum. Terdapat 30 *record* data provinsi yang masuk dalam *cluster* ketiga, dan menjadi *cluster* yang memiliki jumlah objek terbanyak. Anggota *cluster* ini diantaranya adalah semua provinsi di wilayah pulau Sumatera, provinsi DIY, provinsi Banten, provinsi Bali, provinsi NTT, provinsi NTB, semua provinsi di wilayah pulau Kalimantan, semua provinsi di wilayah pulau Sulawesi, provinsi Maluku dan Maluku Utara, serta provinsi Papua dan Papua Barat.

3.4 Tahapan Pengujian

Pada proses ini metode klasterisasi dengan algoritma K-Means diterapkan untuk pembentukan kelompok *cluster* dengan keakurasian yang tepat. Untuk acuan dalam menentukan algoritma tersebut benar. Selain dilakukan perhitungan manual maka pengujian metode juga perlu dilaukan dan dalam penelitian ini penulis menggunakan pengujian perhitungan dengan bantuan aplikasi RapidMiner Studio.



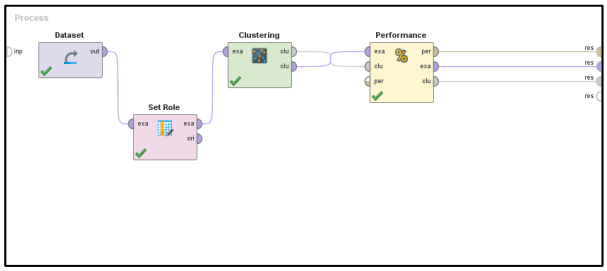
Gambar 4.1 Tampilan aplikasi RapidMiner Studio

Jumlah dataset yang akan dimodelkan dalam pengujian ini merupakan dataset yang juga digunakan pada proses perhitungan manual seperti pada tabel 4.1, yang sudah di *pra-processing* pada tahapan sebelumnya. Dataset yang sudah melalui tahapan *pra-processing* perlu di *import* terlebih dahulu ke dalam aplikasi RapidMiner Studio. Proses *import* data ke dalam aplikasi RapidMiner Studio dimaksudkan agar pembacaan dataset dapat optimal dan mampu digunakan dalam melakukan pemodelan yang sudah ditentukan.

Row No.	Nama Prov	Kasus Konfirmasi	Kasus Sembuh
1	Aceh	43315	38906
2	Sumatera Ut...	153990	146280
3	Sumatera Ba...	103181	96188
4	Riau	150006	142874
5	Jambi	37876	35966
6	Sumatera Sel...	80077	74259
7	Bengkulu	28978	26866
8	Kepulauan B...	64915	61441
9	Lampung	73308	57156
10	Kepulauan Ri...	69337	63680
11	DKI Jakarta	1231444	1202905
12	Jawa Barat	1091786	1023684
13	Jawa Tengah	620629	565091
14	Daerah Istim...	218190	184920
15	Jawa Timur	571020	535001

Gambar 4.2 Contoh dataset yang digunakan pada aplikasi RapidMiner Studio

Penggunaan beberapa fungsi yang ada pada aplikasi RapidMiner Studio saling dihubungkan sehingga membentuk diagram proses pengujian model data mining dengan metode klasterisasi dan algoritma *K-Means* yang terlihat seperti pada gambar berikut.



Gambar 4.3 Diagram proses algoritma K-Means pada aplikasi RapidMiner Studio

3.5 Evaluasi Model

Setelah melakukan tahapan dalam mencari kelompok klaster melalui metode klasterisasi, pemanfaatan algoritma *K-Means* yang digunakan menghasilkan suatu pengelompokkan *cluster* terhadap masing-masing data yang ada pada dataset. Hasil *cluster model* pada pengujian aplikasi RapidMiner Studio menunjukkan dari 34 data tersebut, terdapat 30 data provinsi masuk ke dalam kelompok objek klaster pertama (C0), kemudian sebanyak 2 data provinsi masuk ke dalam kelompok objek klaster kedua (C1), dan 2 data provinsi masuk ke dalam kelompok objek klaster ketiga (C3), yang dapat dilihat pada gambar berikut.

Cluster Model

```
Cluster 0: 30 items
Cluster 1: 2 items
Cluster 2: 2 items
Total number of items: 34
```

Gambar 4.4 Hasil *cluster* model pada aplikasi RapidMiner Studio

Titik klaster paling optimal untuk masing-masing atribut bagi *Cluster* 0 (C0) ,*Cluster* 1 (C1) dan *Cluster* 2 (C2) seperti yang dapat dilihat pada gambar dibawah ini.

Attribute	cluster_0	cluster_1	cluster_2
Kasus Konfirmasi	82203.633	595824.500	1161615
Kasus Sembuh	76115.767	550046	1113294.500

Gambar 4.5 Hasil optimal *cluster* pada aplikasi RapidMiner Studio

Hasil nilai evaluasi dengan Davies Bouldin Index atau DBI berdasar pada aplikasi RapidMiner Studio yang didapat dari pengujian yang didapat dari hasil uji pada aplikasi RapidMiner Studio meunjukkan angka 0,083 seperti pada gambar 4.6. berikut.

```
PerformanceVector:
Davies Bouldin: 0.083
```

Gambar 4.6 Hasil Davies-Bouldin Index pada aplikasi RapidMiner Studio

Pengujian performa terhadap model dan algoritma dilakukan dengan maksud mengetahui hasil perhitungan yang dianalisa dan mengukur metode serta algoritma yang digunakan apakah berfungsi dengan baik atau tidak.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 4044965819.608
Avg. within centroid distance_cluster_0: 4126036389.006
Avg. within centroid distance_cluster_1: 420807622.625
Avg. within centroid distance_cluster_2: 6453065475.625
```

Gambar 4.7 *Performa Clustering* pada aplikasi RapidMiner Studio

3.6 Pembahasan

Setelah melakukan tahapan dalam mencari kelompok klaster melalui metode klasterisasi, pemanfaatan algoritma *K-Means* yang digunakan menghasilkan suatu pengelompokkan klaster terhadap masing-masing data. Hasil penerapan metode klasterisasi dengan algoritma *K-Means* pada dataset kasus Covid-19 provinsi di Indonesia yang digunakan setelah dimodelkan dengan melalui tahapan yang ditentukan mendapatkan posisi *cluster* optimal pada iterasi ke-4, dengan jumlah objek masing-masing pada setiap klaster adalah 2 objek pada klaster pertama, 2 objek pada klaster kedua dan 30 objek pada klaster ketiga.

Berikut merupakan hasil kelompok *cluster* yang masuk pada *cluster* pertama, *cluster* kedua dan *cluster* ketiga untuk setiap masing-masing provinsi yang dapat dilihat pada tabel 4.15.

Tabel 4.15 Kelompok objek data pada setiap *cluste*

Nama Prov	*Cluster*	Kategori
Aceh	C3	Prioritas Umum
Sumatera Utara	C3	Prioritas Umum
Sumatera Barat	C3	Prioritas Umum
Riau	C3	Prioritas Umum
Jambi	C3	Prioritas Umum
Sumatera Selatan	C3	Prioritas Umum
Bengkulu	C3	Prioritas Umum
Kepulauan Bangka Belitung	C3	Prioritas Umum
Lampung	C3	Prioritas Umum
Kepulauan Riau	C3	Prioritas Umum
DKI Jakarta	C1	Prioritas Utama
Jawa Barat	C1	Prioritas Utama
Jawa Tengah	C2	Prioritas Tinggi
Daerah Istimewa Yogyakarta	C3	Prioritas Umum
Jawa Timur	C2	Prioritas Tinggi
Banten	C3	Prioritas Umum
Bali	C3	Prioritas Umum
Nusa Tenggara Barat	C3	Prioritas Umum
Nusa Tenggara Timur	C3	Prioritas Umum
Kalimantan Barat	C3	Prioritas Umum
Kalimantan Tengah	C3	Prioritas Umum
Kalimantan Selatan	C3	Prioritas Umum
Kalimantan Timur	C3	Prioritas Umum
Kalimantan Utara	C3	Prioritas Umum
Sulawesi Utara	C3	Prioritas Umum
Sulawesi Tengah	C3	Prioritas Umum
Sulawesi Selatan	C3	Prioritas Umum
Sulawesi Tenggara	C3	Prioritas Umum
Gorontalo	C3	Prioritas Umum
Sulawesi Barat	C3	Prioritas Umum
Maluku	C3	Prioritas Umum
Maluku Utara	C3	Prioritas Umum
Papua	C3	Prioritas Umum
Papua Barat	C3	Prioritas Umum

Hasil pembentukan *cluster model* pada pengujian aplikasi RapidMiner Studio juga memiliki kemiripan dengan proses perhitungan manual, dimana dengan melakukan *set-up* klaster sebanyak 3 klaster maka terbentuk kelompok objek untuk masing-masing klaster dengan angggota kelompok sebanyak 30 objek masuk pada kelompok *Cluster* 0 (C0), 2 objek masuk pada kelompok *Cluster* 1 (C1), dan terakhir ada 2 objek masuk pada kelompok *Cluster* 2 (C2).

Row No.	Nama Prov	label	Kasus Konfi...	Kasus Sem...
1	Aceh	cluster_0	43315	38906
2	Sumatera Ut...	cluster_0	153990	146260
3	Sumatera Ba...	cluster_0	103181	96188
4	Riau	cluster_0	150906	142874
5	Jambi	cluster_0	37876	35996
6	Sumatera Sel...	cluster_0	80077	74259
7	Bengkulu	cluster_0	28876	26856
8	Kepulauan B...	cluster_0	94915	91441
9	Lampung	cluster_0	73308	57156
10	Kepulauan Ri...	cluster_0	69337	63690
11	DKI Jakarta	cluster_2	1231444	1202905
12	Jawa Barat	cluster_2	1091786	1023684
13	Jawa Tengah	cluster_1	620629	565091
14	Daerah Istim...	cluster_0	218190	184920
15	Jawa Timur	cluster_1	571020	535061
16	Banten	cluster_0	286512	269392
17	Bali	cluster_0	155925	150213
18	Nusa Tengga...	cluster_0	35949	34747
19	Nusa Tengga...	cluster_0	91684	82541
20	Kalimantan B...	cluster_0	63059	58568
21	Kalimantan T...	cluster_0	55158	52263
22	Kalimantan S...	cluster_0	83984	80879
23	Kalimantan T...	cluster_0	205128	195663
24	Kalimantan U...	cluster_0	44979	41339
25	Sulawesi Utara	cluster_0	51312	49232
26	Sulawesi Ten...	cluster_0	59982	54828
27	Sulawesi Sel...	cluster_0	142746	136407
28	Sulawesi Ten...	cluster_0	25455	24464
29	Gorontalo	cluster_0	13803	12853
30	Sulawesi Barat	cluster_0	15381	14336
31	Maluku	cluster_0	18517	18137
32	Maluku Utara	cluster_0	14348	13220
33	Papua	cluster_0	47985	35578
34	Papua Barat	cluster_0	31131	30325

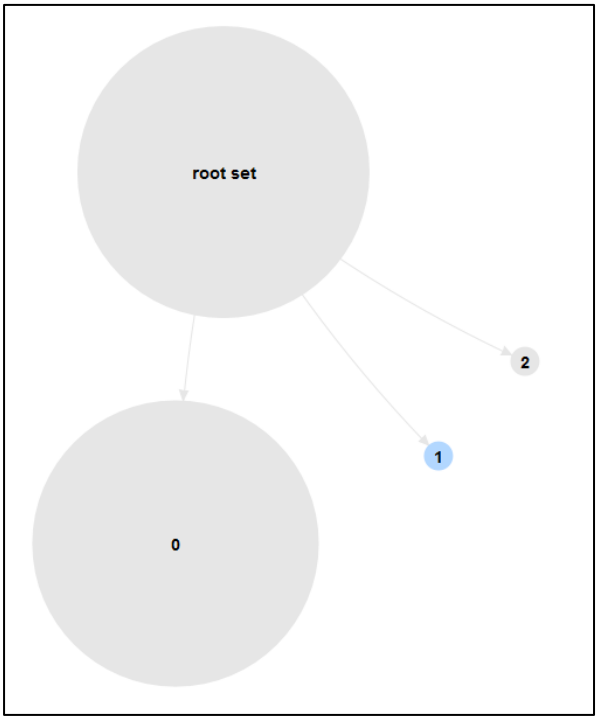
Gambar 4.8 *Cluster* objek pada aplikasi RapidMiner Studio

Berikut merupakan perbandingan objek klaster pada perhitungan manual yang dilakukan dengan pengujian menggunakan aplikasi RapidMiner Studio pada tabel 4.16.

Tabel 4.16 Perbandingan objek data dari hasil perhitungan manual dengan aplikasi RapidMiner Studio

Nama Prov	*Cluster (manual)*	*Cluster (RapidMiner)*
Aceh	C3	*Cluster_0*
Sumatera Utara	C3	*Cluster_0*
Sumatera Barat	C3	*Cluster_0*
Riau	C3	*Cluster_0*
Jambi	C3	*Cluster_0*
Sumatera Selatan	C3	*Cluster_0*
Bengkulu	C3	*Cluster_0*
Kepulauan Bangka Belitung	C3	*Cluster_0*
Lampung	C3	*Cluster_0*
Kepulauan Riau	C3	*Cluster_0*
DKI Jakarta	C1	*Cluster_2*
Jawa Barat	C1	*Cluster_2*
Jawa Tengah	C2	*Cluster_1*
Daerah Istimewa Yogyakarta	C3	*Cluster_0*
Jawa Timur	C2	*Cluster_1*
Banten	C3	*Cluster_0*
Bali	C3	*Cluster_0*
Nusa Tenggara Barat	C3	*Cluster_0*
Nusa Tenggara Timur	C3	*Cluster_0*
Kalimantan Barat	C3	*Cluster_0*
Kalimantan Tengah	C3	*Cluster_0*
Kalimantan Selatan	C3	*Cluster_0*
Kalimantan Timur	C3	*Cluster_0*
Kalimantan Utara	C3	*Cluster_0*
Sulawesi Utara	C3	*Cluster_0*
Sulawesi Tengah	C3	*Cluster_0*
Sulawesi Selatan	C3	*Cluster_0*
Sulawesi Tenggara	C3	*Cluster_0*
Gorontalo	C3	*Cluster_0*
Sulawesi Barat	C3	*Cluster_0*
Maluku	C3	*Cluster_0*
Maluku Utara	C3	*Cluster_0*
Papua	C3	*Cluster_0*
Papua Barat	C3	*Cluster_0*

Sehingga setiap objek klaster memiliki kemiripan hasil antara perhitungan manual dengan pengujian aplikasi RapidMiner. Adapaun gambaran umum kelompok klaster dari masing-masing klaster 0, 1 dan 2 dapat dilihat pada pohon klasterisasi seperti berikut ini. Gambar pohon klasterisasi ini menunjukkan bahwa sebagian besar data yang digunakan terkelompok pada klaster 0, diikuti kemudian oleh klaster 1 dan klaster 2.



Gambar 4.9 *Cluster Tree* yang terbentuk pada aplikasi RapidMiner Studio

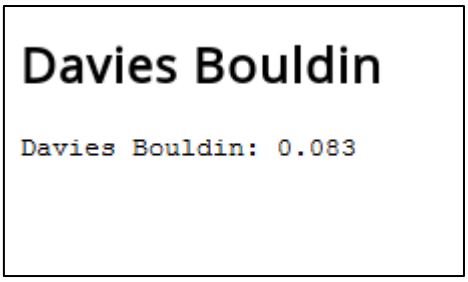
Titik klaster optimal yang terbentuk dari masing-masing atribut yang digunakan yakni atribut kasus konfirmasi dan kasus konfirmasi sembuh jika disandingkan dengan hasil perhitungan model algoritma *K-Means* yang dilakukan secara manual juga memiliki kemiripan seperti pada tabel berikut.

Tabel 4.17 Nilai klaster optimal pada perhitungan algoritma K-Means secara manual

Titik Pusat Awal	Manual		RapidMiner	
	Kasus Konfirmasi	Kasus Sembuh	Kasus Konfirmasi	Kasus Sembuh
Klaster 1	1.161.615	1.113.295	1.161.615	1.113.294,5
Klaster 2	595.825	550.046	595.824,5	550.046
Klaster 3	82.204	76.116	82.203,6	76.115,8

Evaluasi model dengan melihat nilai *Davies Bouldin Index* atau DBI yang dihasilkan pada aplikasi RapidMiner Studio yang didapat dari

pengujian yang meunjukkan angka 0,083 seperti pada gambar 4.10.



Gambar 4.10 Nilai hasil *Davies Bouldin Index* pada aplikasi RapidMiner Studio

4. Kesimpulan

Berdasarkan hasil yang penelitian yang dilakukan oleh penulis, kesimpulan yang didapat yaitu sebagai berikut :

1. Analisis kebutuhan penyelenggaraan vaksinasi berdasar tingkat wilayah prioritas dengan memanfaatkan data sebaran kasus COVID-19 di setiap provinsi di Indonesia dapat menghasilkan suatu model klaster dengan menggunakan metode klasterisasi sesuai dengan tahapan yang ditentukan.
2. Algoritma K-Means digunakan untuk menunjukkan sebuah pengetahuan dan perspektif baru, yakni dengan mengelompokkan provinsi di Indonesia berdasarkan 3 klaster yang diusulkan, yakni klaster 1 merupakan tingkat prioritas tinggi yang terdiri dari 2 provinsi, klaster 2 adalah prioritas utama yang terdiri dari 2 provinsi, dan yang ketiga adalah klaster 3 yakni prioritas umum yang terdiri dari 30 provinsi di Indonesia.
3. Pengujian menggunakan aplikasi RapidMiner Studio juga menghasilkan wawasan yang relevan terhadap model yang ditentukan saat perhitungan manual yang dimana memiliki nilai evaluasi *Davies-Bouldin Index* untuk model algoritma tersebut sebesar 0,083.

Ucapan Terima Kasih

Puji syukur penulis panjatkan kehadiran Allah SWT, Tuhan yang maha pengasih dan maha penyayang yang telah melimpahkan segala rahmat dan hidayah-Nya, sehingga tersusunlah skripsi yang berjudul "ANALISA DATA MINING UNTUK KLASTERISASI VAKSIN COVID-19 MENGGUNAKAN ALGORITMA *K-MEANS*" Dapat penulis selesaikan sesuai dengan rencana karena dukungan dari berbagai pihak yang tidak ternilai besarnya.

Pada kesempatan ini penulis mengucapkan rasa teima kasih yang sedalam-dalamnya kepada:

1. a Bapak Hamzah Muhammad Mardi Putra, S.K.M., M.M., D.B.A., sebagai Rektor Universitas Pelita Bangsa.
2. Ibu Putri Anggun Sari, S.Pt., M.Si., sebagai Dekan Fakultas Teknik, Universitas Pelita Bangsa.
3. Bapak Wahyu Hadikristanto, S.Kom., M.Kom., sebagai ketua Program Studi Teknik Informatika S-1.
4. Bapak Ahmad Turmudy Zy, M.Kom., sebagai Dosen Pembimbing I.
5. Ibu Edora, S.Pd., M.Pd., sebagai Dosen Pembimbing II
6. Kedua orang tua saya yang telah membantu memberikan semangat, doa, kasih harapan dan dorongan moral dan spiritual kepada penulis dan menyelesaikan tugas akhir skripsi.
7. Semua bapak/ibu dosen, keluarga, teman, yang telah membantu penulis dalam penyusunan skripsi ini sampai selesai yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam penulisan tugas akhir ini jauh dari kata sempurna, untuk itu penulis sangat mengharapkan dukungan dari para pembaca untuk memberikan kritik dan saran yang membangun guna perbaikan dimasa yang akan datang.

Referensi

[1] A. Baita, Y. Pristyanto, N. Cahyono, P. Covid-, K. N. N. Akurasi, and K. Kunci, "ANALISIS SENTIMEN MENGENAI VAKSIN SINOVAC MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN K-NEAREST NEIGHBOR (KNN) Abstraksi Keywords :," vol. 4, no. 2, pp. 42–46, 2021.

[2] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J. ISSN*, vol. 8, no. 2, pp. 2655–9129, 2020.

[3] A. Harun and D. P. Ananda, "Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes

dan Decission Tree," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. April, pp. 58–63, 2021.

[4] O. Purwaningrum, Y. Y. Putra, and A. A. Arifiyanti, "Penentuan Kelompok Status Gizi Balita dengan Menggunakan Metode K-Means," *J. Ilm. Teknol. Inf. Asia*, vol. 15, no. 2, pp. 129–136, 2021.

[5] P. Penyebaran *et al.*, "Analisis algoritma k-medoids clustering dalam pengelompokan penyebaran covid-19 di indonesia," vol. 4, no. 1, pp. 166–173, 2020.

[6] R. E. Reigal, J. L. Pastrana-Brincones, S. L. González-Ruiz, A. Hernández-Mendo, J. P. Morillo-Baro, and V. Morales-Sánchez, "Use of Data Mining to Determine Usage Patterns of an Online Evaluation Platform During the COVID-19 Pandemic," *Front. Psychol.*, vol. 11, no. September, pp. 1–16, 2020, doi: 10.3389/fpsyg.2020.588843.

[7] A. P. Windarto, U. Indriani, M. R. Raharjo, and L. S. Dewi, "Bagian 1: Kombinasi Metode *Cluster*ing dan Klasifikasi (Kasus Pandemi Covid-19 di Indonesia)," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 855, 2020, doi: 10.30865/mib.v4i3.2312.

[8] I. Romli, S. Prameswari R, and A. Z. Kamalia, "Sentiment Analysis about Large-Scale Social Restrictions in Social Media Twitter Using Algoritm K-Nearest Neighbor," *J. Online Inform.*, vol. 6, no. 1, p. 96, 2021, doi: 10.15575/join.v6i1.670.

[9] P. D. O. Davies, "Multi-drug resistant tuberculosis," *CPD Infect.*, vol. 3, no. 1, pp. 9–12, 2002.

[10] C. Medel-ramírez, H. Medel-lópez, U. Veracruzana, I. De Investigaciones, and S. Económicos, "Data Article . Data mining for the study of the Epidemic (SARS- CoV-2) COVID-19 : Algorithm for the Authors Corresponding author A bstract Keywords," pp. 1–8.

[11] Suyanto, *Data Mining*. Yogyakarta: Informatika, 2017.

[12] Retno Tri vulandari, *Data Mining*. Yogyakarta: Gava Media, 2017.

[13] O. Villacampa, "(Weka - Thesis) Feature Selection and Classification Methods for Decision Making: A Comparative Analysis," *ProQuest Diss. Theses*, no. 63, p. 188, 2015.

[14] G. widi N. Dicky Nofriansyah, *Algoritma Data Mining Dan pengujian*. Yogyakarta: Cv Budi Utama, 2015.

[15] M. Miftakhul and S. Prihandoko, "Penerapan Algoritma K-Means dan Cure Dalam Menganalisa Pola Perubahan Belanja Dari Retail ke E-Commerce," vol. 7, no. 2, pp. 44–49, 2017.

[16] S. Haryati, A. Sudarsono, and E. (2015) Suryana, "Implementasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.