Jurnal Siliwangi Vol.7. No.1, 2021 **P-ISSN 2477-3891**
Seri Sains dan Teknologi **E-ISSN 2615-4765**

IMPLEMENTASI ALGORITMA K-MEANS DAN C4.5 DALAM MENENTUKAN TINGKAT PENYEBARAN COVID-19 DI INDONESIA

Fadilah Salsabila[1)], Sheila Maulida Intani[2)]

[1,2]Program Studi Informatika Fakultas Teknik Universitas Siliwangi
Jl. Siliwangi No.24, Kahuripan, Kec. Tawang, Tasikmalaya, Jawa Barat 46115
e-mail: 177006006@student.unsil.ac.id[1], 177006012@student.unsil.ac.id[2]

Abstrak

COVID-19 merupakan penyakit menular yang disebabkan oleh virus corona yang sedang melanda dunia, termasuk Indonesia. Saat ini, penyebaran kasus *Covid-19* cukup cepat dan sangat berdampak negatif terhadap semua bidang. Indonesia memiliki wilayah yang luas, sehingga penelitian ini bertujuan untuk mengelompokkan tingkat penyebaran kasus *Covid-19* berdasarkan provinsi di Indonesia. Pengelompokkan menggunakan kombinasi algoritma klasterisasi *K-Means* dengan algoritma klasifikasi C4.5. Algoritma *K-Means* berfungsi untuk melakukan pengelompokkan data ke dalam klaster wilayah di Indonesia berdasarkan provinsi. Hasil dari pengelompokkan digunakan algoritma C4.5 untuk melihat aturan berupa pohon keputusan. Label pengelompokkan yang digunakan sebanyak 4 klaster yaitu klaster darurat (cluster_0 = hitam), klaster tinggi (cluster_1 = zona merah), klaster sedang (cluster_2 = zona kuning), dan klaster rendah (cluster_3 = zona hijau). Penentuan jumlah klaster (k) ditentukan dengan menggunakan parameter DBI (Davies Bouldin Index) untuk mengoptimalkan hasil klaster yang diperoleh, dimana untuk k=4 ini memiliki nilai DBI sebesar 0.110. Hasil pengelompokkan yang diperoleh terdapat 1 provinsi yang berada di klaster darurat yaitu DKI Jakarta, 3 provinsi yang berada di klaster tinggi yaitu Jawa Barat, Jawa Tengah, dan Jawa Timur, 8 provinsi yang berada di klaster sedang yaitu Sumatera Utara, Sumatera Barat, Riau, Kalimantan Timur, Kalimantan Selatan, Banten, Bali serta Sulawesi Selatan, dan sisanya sebanyak 22 provinsi berada di klaster rendah. Hasil dari kombinasi algoritma tersebut dapat digunakan untuk kebutuhan tertentu dan memberikan pengetahuan baru yaitu informasi mengenai pemetaan berupa klaster terhadap jumlah persebaran kasus *Covid-19* di Indonesia.

Kata Kunci : Algoritma C4.5, *Covid-19*, *Data Mining, K-Means*.

Abstract

COVID-19 is an infectious disease caused by the corona virus that is sweeping the world, including Indonesia. Currently, the spread of Covid-19 cases is quite fast and has a very negative impact on all fields. Indonesia has a large area, so this study aims to classify the level of spread of Covid-19 cases by province in Indonesia. The grouping uses a combination of the K-Means clustering algorithm with the C4.5 classification algorithm. The K-Means algorithm functions to group data into regional clusters in Indonesia by province. The results of the grouping used the C4.5 algorithm to see the rules in the form of a decision tree. The grouping labels used were 4 clusters, namely emergency cluster (cluster_0 = black), high cluster (cluster_1 = red zone), medium cluster (cluster_2 = yellow zone), and low cluster (cluster_3 = green zone). The determination of the number of clusters (k) is determined by using the DBI (Davies Bouldin Index) parameter to optimize the results of the obtained clusters, where for k=4 it has a DBI value of 0.110. The grouping results obtained are 1 province in the emergency cluster, namely DKI Jakarta, 3 provinces in the high cluster, namely West Java, Central Java, and East Java, 8 provinces in the medium cluster, namely North Sumatra, West Sumatra, Riau, East Kalimantan, South Kalimantan, Banten, Bali and South Sulawesi, and the remaining 22 provinces are in the low cluster. The results of the combination of these algorithms can be used for certain needs and provide new knowledge, namely information on mapping in the form of clusters on the number of Covid-19 cases in Indonesia.

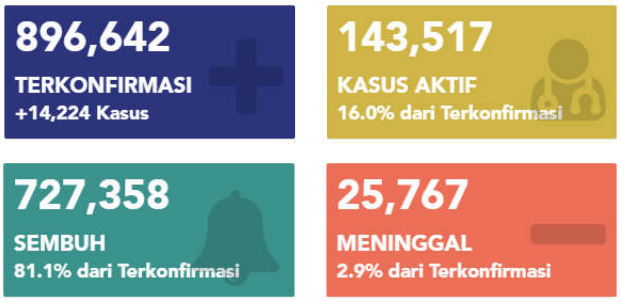
Keywords: C4.5 Algorithm, Covid-19, Data Mining, K-Means.

I. PENDAHULUAN

COVID-19 merupakan penyakit menular yang disebabkan oleh coronavirus jenis baru yaitu SARS-CoV-2 dan ditandai dengan adanya gejala seperti demam, batuk dan sesak nafas [1]. Virus corona ini sangat berbahaya karena dapat menyebabkan gejala mematikan seperti sesak nafas, dada sakit dan kesulitan dalam berbicara sampai berujung pada kematian [2]. Virus ini tidak hanya dapat menyerang manusia tetapi virus ini juga dapat menyerang hewan terutama hewan peliharaan seperti anjing dan kucing [1]. Indonesia memiliki wilayah yang sangat luas, sehingga perlu dilakukan pengelompokkan penyebaran kasus *Covid-19* berdasarkan wilayah di Indonesia. Pengelompokkan ini akan menghasilkan titik-titik pusat penyebaran kasus pandemi *Covid-19* di Indonesia [3].

Data mining merupakan sebuah alat atau proses penambangan informasi penting dari suatu data berukuran besar dengan mencari pola atau informasi menarik melalui penguraian penemuan pengetahuan di dalam *database* dengan menggunakan teknik tertentu [4]. Teknik dalam data mining diantaranya *clustering* dan klasifikasi [5]. Algoritma *clustering* dan klasifikasi yang saat ini banyak digunakan diantaranya *K-Means* dan C4.5 [6]. Algoritma *K-Means* digunakan untuk pengelompokkan iteratif dengan cara melakukan partisi set data kedalam sejumlah cluster (k) yang sudah ditetapkan diawal [7]. Sedangkan Algoritma C4.5 digunakan untuk membentuk pohon keputusan (*decision tree*) yang dapat digunakan untuk memprediksi sebuah keputusan dengan menerapkan serangkaian aturan keputusan [8].

Berdasarkan hal tersebut menggabungkan kedua algoritma dalam suatu penelitian menjadi hal yang sangat menarik untuk dilakukan. Kasus yang digunakan untuk menguji gabungan kedua algoritma tersebut adalah jumlah persebaran *Covid-19* di Indonesia yang bersumber dari Data Kementerian Kesehatan RI yang dapat diakses melalui tautan covid19.go.id. Gambar 1 merupakan data keseluruhan peta persebaran kasus *Covid-19* di Indonesia pada tanggal 10 Januari 2021.



Gambar 1. Data Penderita Covid 19 (10 Jan 2021)

Tujuan dari penelitian ini untuk membangun model kombinasi algoritma *clustering* dan klasifikasi pada kasus jumlah persebaran pandemi *Covid-19* di Indonesia dan menguji model kombinasi tersebut.

II. BAHAN DAN METODE/METODOLOGI

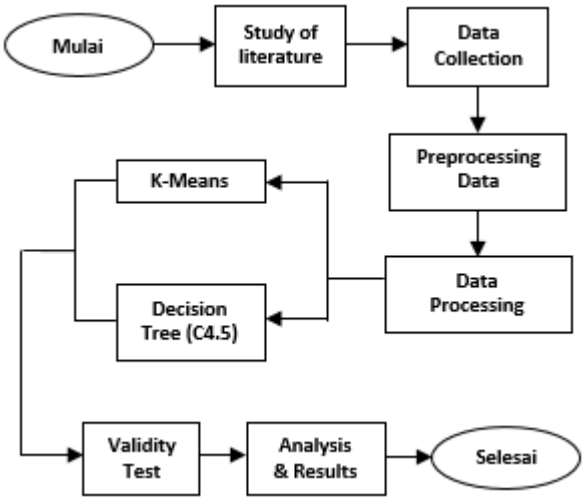
A. Tahap Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari Kementerian Kesehatan RI melalui covid19.go.id. Data yang digunakan adalah data jumlah persebaran kasus pandemi *Covid-19* di Indonesia pada tanggal 10 januari 2021 yang terdiri dari 34 *record* seperti yang ditunjukkan pada tabel 1.

Tabel 1. Data Jumlah Persebaran Kasus Covid-19

No.	Provinsi	Jumlah Positif	Jumlah Meninggal	Jumlah Sembuh
1	Aceh	8909	363	7298
2	Sumatera Utara	19027	697	16259
3	Sumatera Selatan	12522	618	10095
4	Sumatera Barat	24509	543	18259
5	Bengkulu	4062	128	3381
6	Riau	26332	618	24301
7	Kepulauan Riau	7306	178	6564
8	Jambi	3601	58	2609
9	Lampung	7213	351	5013
10	Bangka Belitung	3005	52	2313
11	Kalimantan Barat	3348	28	3034
12	Kalimantan Timur	30511	804	24740
13	Kalimantan Selatan	16079	598	14378
14	Kalimantan Tengah	10446	200	5044
15	Kalimantan Utara	4843	64	2292
16	Banten	20513	413	10794
17	DKI Jakarta	206122	3485	184438
18	Jawa Barat	97570	1219	81774
19	Jawa Tengah	91715	3742	61525
20	DI Yogyakarta	14929	325	9892
21	Jawa Timur	92613	6441	79667
22	Bali	19232	552	12477
23	Nusa Tenggara Timur	2526	61	1501
24	Nusa Tenggara Barat	6136	263	3909
25	Gorontalo	3969	106	3607
26	Sulawesi Barat	2285	48	1654
27	Sulawesi Tengah	4649	124	2515
28	Sulawesi Utara	10470	337	7516
29	Sulawesi Tenggara	8400	162	7043
30	Sulawesi Selatan	36513	648	30929
31	Maluku Utara	2940	92	2486
32	Maluku	5881	74	4791
33	Papua Barat	6188	103	5663
34	Papua	13662	155	8078

Terdapat beberapa tahap yang dilakukan pada penelitian inim diantaranya : Studi literature, data collection, preprocessing data, data processing (K-Means & Decission Tree C4.5), Validity Test, Analysis & Result. Secara umum tahapan penelitian yang dilakukan ditampilkan pada gambar 2.



Gambar 2. Metodologi Penelitian

III. HASIL DAN PEMBAHASAN

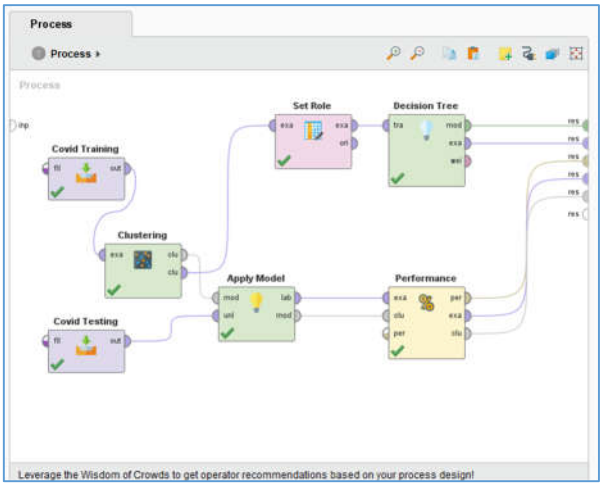
Pada tahap implementasi digunakan RapidMiner Studio 9.8. Data yang diambil adalah 34 data yang sudah melewati proses *Pre-Processing*, meliputi data jumlah kasus positif, jumlah meninggal, dan jumlah sembuh. Setelah data terkumpul, dilakukan analisis data yang sesuai dengan kebutuhan sistem, yaitu melakukan klastering dengan algoritma K-Means dan klasifikasi dengan algoritma C4.5. Gabungan dari algoritma *clustering* dan klasifikasi memiliki peran masing-masing. Algoritma *clustering* yaitu *K-Means* berperan melakukan pemetaan berupa klaster dan hasilnya kemudian diproses menggunakan algoritma klasifikasi yaitu C4.5 untuk melihat nilai aturan berupa pohon keputusan. Sebelum melakukan pemetaan, penentuan jumlah klaster dilakukan dengan menggunakan parameter DBI. Berikut grafik perbandingan jumlah klaster (k) dengan menggunakan *Davies Bouldin Index.*

Tabel 2. Hasil Perbandingan DBI Klaster

K-Means	
Klaster	Nilai DBI
K = 2	0.158
K = 3	0.176
K = 4	**0.110**
K = 5	0.116
K = 6	0.177

Berdasarkan perbandingan nilai Davies-Bouldin Index table 2 jumlah klaster (k=4) menjadi klaster terbaik karena memiliki nilai terkecil yakni 0.110 (paling optimal). Berdasarkan hal tersebut jumlah klaster yang akan digunakan adalah 4 label yakni Klaster Darurat (Zona Hitam), Klaster Tinggi (Zona Merah), Klaster Sedang (Zona Kuning), dan Klaster Rendah (Zona Hijau).

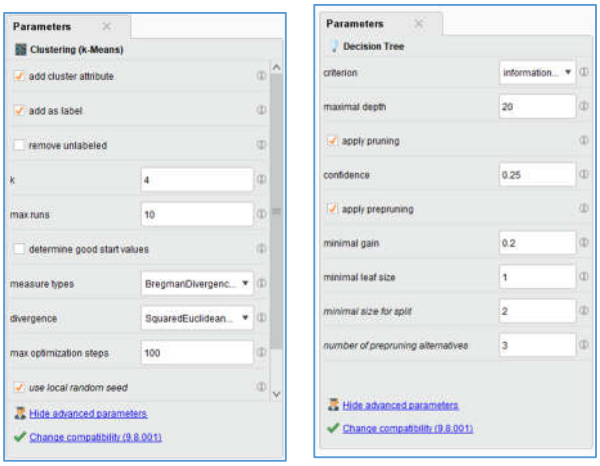
Model penggabungan algoritma *clustering* dan klasifikasi dengan menggunakan software RapidMiner Studio 9.8 ditampilkan pada gambar 3.



Gambar 3. *View Process* pada RapidMiner Studio

Pada gambar 3 ditampilkan beberapa parameter digunakan untuk menghasilkan output yang sesuai dengan tujuan, diantaranya: *set role*, *apply model*, *performance*, *clustering* (k-means) dan *decision tree*(C4.5). Setiap parameter memiliki tugas masing-masing. Read excel (covid training) yang digunakan untuk input data (tabel 1) diproses menuju *clustering* (k-means) dan *set role*. *Set role* memeriksa kebenaran label klaster yang digunakan pada data. Apabila sudah benar, *apply model* akan menampilkan visual data berupa grafik dan tabel dengan menggunakan data yang bersumber dari *read excel* (*covid testing* = *covid training*). Data tersebut kemudian diproses menggunakan algoritma C4.5 (*decision tree*) dan yang terakhir akan diukur *performance* dari hasil yang diperoleh. Gambar 4 merupakan tampilan pengaturan parameter pada setiap metode atau algoritma.

P-ISSN 2477-3891
E-ISSN 2615-4765



Gambar 4. Pengaturan parameter pada setiap metode (K-means dan C4.5)

Hasil pemetaan berupa klaster pada jumlah persebaran kasus pandemi *Covid-19* di Indonesia akan ditampilkan seperti pada gambar 5.

Gambar 5. Hasil pemetaan klaster

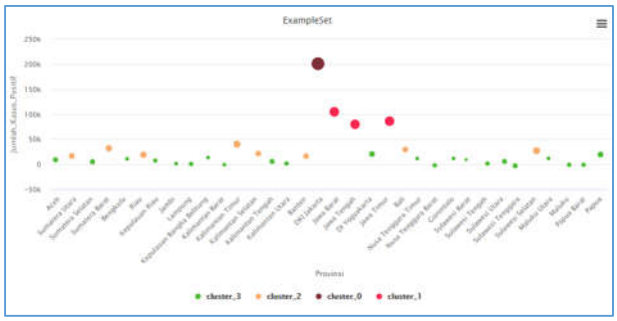
Pada gambar 5 ditampilkan hasil dari pemetaan atau pengelompokkan, klaster darurat (zona hitam) terdapat 1 provinsi (cluster_0), klaster tinggi (zona merah) terdapat 3 provinsi (cluster_1) , klaster sedang (zona kuning) terdapat 8 provinsi (cluster_2) dan klaster rendah (zona hijau) terdapat 22 provinsi (cluster_3). Berikut hasil lengkap pemetaan klaster terhadap jumlah persebaran pandemi *Covid-19* di Indonesia:

a. Klaster Darurat (zona hitam) adalah DKI Jakarta.
b. Klaster Tinggi (zona merah) adalah Jawa Barat, Jawa Tengah, dan Jawa Timur.
c. Klaster Sedang (zona kuning) adalah Sumatera Utara, Sumatera Barat, Riau, Kalimantan Timur, Kalimantan Selatan, Banten, Bali, dan Sulawesi Selatan.
d. Klaster Rendah (zona hijau) adalah Aceh, Sumatera Selatan, Bengkulu, Kepulauan Riau, Jambi, Lampung, Kepulauan Bangka Belitung, Kalimantan Barat, Kalimantan Tengah, Kalimantan Utara, DI Yogyakarta, Nusa Tenggara Timur, Nusa Tenggara Barat, Gorontalo, Sulawesi Barat, Sulawesi Tengah, Sulawesi Utara, Sulawesi Tenggara, Maluku Utara, Maluku, Papua Barat, dan Papua.

Dalam penentuan klaster (cluster_0, 1, 2, dan 3) juga didasarkan atas hasil centroid akhir yang ditunjukkan oleh gambar 6.

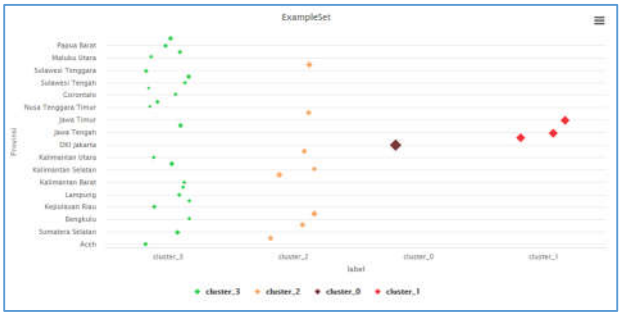
Gambar 6. Hasil centroid akhir (k-means)

Hasil dari pemetaan klaster dapat di visualisasikan dengan diagram plot scatter seperti yang terlihat pada gambar 7.



Gambar 7. Scatter plot berdasarkan jumlah kasus positif untuk tiap provinsi dengan klaster

P-ISSN 2477-3891
E-ISSN 2615-4765

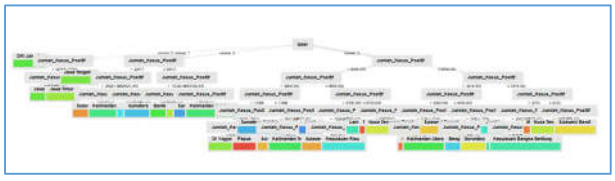


Gambar 8. Scatter plot berdasarkan cluster untuk tiap provinsi

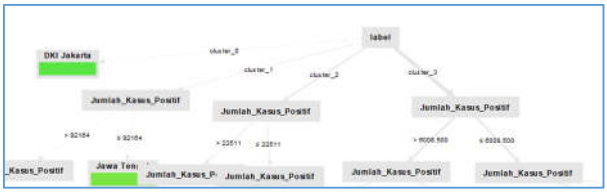
	Provinsi	Jumlah_Kasus_Positif	Jumlah_Meninggal	Jumlah_Sembuh	label
2	Aceh	8909,0	363,0	7298,0	cluster_3
3	Sumatera Utara	19027,0	697,0	16259,0	cluster_2
4	Sumatera Selatan	12522,0	618,0	10095,0	cluster_3
5	Sumatera Barat	24509,0	543,0	18259,0	cluster_2
6	Bengkulu	4062,0	128,0	3381,0	cluster_3
7	Riau	26332,0	618,0	24301,0	cluster_2
8	Kepulauan Riau	7306,0	178,0	6564,0	cluster_3
9	Jambi	3601,0	58,0	2609,0	cluster_3
10	Lampung	7213,0	351,0	5013,0	cluster_3
11	Kepulauan Bangka Belitu	3005,0	52,0	2313,0	cluster_3
12	Kalimantan Barat	3348,0	28,0	3034,0	cluster_3
13	Kalimantan Timur	30511,0	804,0	24740,0	cluster_2
14	Kalimantan Selatan	16079,0	598,0	14378,0	cluster_2
15	Kalimantan Tengah	10446,0	200,0	5044,0	cluster_3
16	Kalimantan Utara	4843,0	64,0	2292,0	cluster_3
17	Banten	20513,0	413,0	10794,0	cluster_2
18	DKI Jakarta	206122,0	3485,0	184438,0	cluster_0
19	Jawa Barat	97570,0	1219,0	81774,0	cluster_1
20	Jawa Tengah	91715,0	3742,0	61525,0	cluster_1
21	DI Yogyakarta	14929,0	325,0	9892,0	cluster_3
22	Jawa Timur	92613,0	6441,0	79667,0	cluster_1
23	Bali	19232,0	552,0	12477,0	cluster_2
24	Nusa Tenggara Timur	2526,0	61,0	1501,0	cluster_3
25	Nusa Tenggara Barat	6136,0	263,0	3909,0	cluster_3
26	Gorontalo	3969,0	106,0	3607,0	cluster_3
27	Sulawesi Barat	2285,0	48,0	1654,0	cluster_3
28	Sulawesi Tengah	4649,0	124,0	2515,0	cluster_3
29	Sulawesi Utara	10470,0	337,0	7516,0	cluster_3
30	Sulawesi Tenggara	8400,0	162,0	7043,0	cluster_3
31	Sulawesi Selatan	36513,0	648,0	30929,0	cluster_2
32	Maluku Utara	2940,0	92,0	2486,0	cluster_3
33	Maluku	5881,0	74,0	4791,0	cluster_3
34	Papua Barat	6188,0	103,0	5663,0	cluster_3
35	Papua	13662,0	155,0	8078,0	cluster_3

Gambar 9. Hasil Data Export Pemetaan Klaster

Pada gambar 9, hasil pemetaan klaster dapat dikonversi ke format lainnya sehingga hasil dari pemetaan tersebut dapat digunakan sesuai dengan kebutuhan. Selanjutnya hasil pemetaan tersebut diproses dengan menggunakan algoritma C4.5 (*decision tree*) untuk melihat informasi berupa aturan pohon keputusan. Berikut ini hasil dari *decision tree* atau pohon keputusan yang diperoleh dari bantuan software RapidMiner Studio 9.8 seperti yang terlihat pada gambar 10



Gambar 10. Hasil pohon keputusan / decision tree secara keseluruhan



Gambar 11. Hasil pohon keputusan / decision tree bagian utama

Dari pohon keputusan dapat diambil suatu informasi berdasarkan klaster bahwa label cluster_0 (klaster darurat (zona hitam)) ditetapkan hanya 1 provinsi yaitu DKI Jakarta saja. Untuk cluster_1 (klaster tinggi (zona merah)) terjadi jika jumlah kasus positif berada diantara nilai 92164. Untuk cluster_2 (klaster sedang (zona kuning)) terjadi jika jumlah kasus positif berada diantara nilai 22511. Untuk cluster_2 (klaster rendah (zona hijau) terjadi jika jumlah kasus positif berada diantara nilai 6008,5.

Selain itu untuk hasil pengujian performance dilakukan dengan cara menggunakan operator cluster distance performance. Operator ini digunakan untuk mengevaluasi kinerja metode pengelompokan berbasis centroid. Operator ini memberikan daftar nilai kriteria kinerja berdasarkan centroid. Cluster distance performance yang dimaksud penggunaan Davies Bouldin Index. Pada penelitian ini menghasilkan nilai DBI yang cukup optimal yaitu 0.110 seperti yang terlihat pada gambar dibawah ini:

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 13631665.334
Avg. within centroid distance_cluster_0: 0.000
Avg. within centroid distance_cluster_1: 31265640.963
Avg. within centroid distance_cluster_2: 27905037.323
Avg. within centroid distance_cluster_3: 6656336.358
Davies Bouldin: 0.110
```

Gambar 12. Hasil Uji Performance

IV. KESIMPULAN

Penggabungan algoritma *clustering* dan klasifikasi dapat diterapkan pada kasus persebaran pandemi *Covid-19* di Indonesia dengan menggunakan bantuan software RapidMiner Studio 9.8. Dengan menggunakan penggabungan kedua algoritma tersebut untuk hasil pemetaan wilayah berupa klaster diperoleh 4 klaster yaitu klaster darurat (zona hitam) terdiri 1 provinsi (cluster_0), klaster tinggi (zona merah) terdiri 3 provinsi (cluster_1) , klaster sedang (zona kuning) terdiri 8 provinsi (cluster_2) dan klaster rendah (zona hijau) terdiri 22 provinsi (cluster_3). Nilai yang diperoleh dari pohon keputusan untuk Untuk cluster_1 terjadi jika jumlah kasus positif berada diantara nilai 92164. Untuk

cluster_2 terjadi jika jumlah kasus positif berada diantara nilai 22.511. Untuk cluster_2 terjadi jika jumlah kasus positif berada diantara nilai 6.008,5. Dan terakhir untuk cluster_0 yang merupakan klaster dengan resiko sangat tinggi yaitu DKI Jakarta memiliki jumlah kasus positif 206.122. Dengan data tersebut dapat digunakan untuk kebutuhan tertentu misalnya dapat membantu pemerintah dalam membuat suatu keputusan atau kebijakan.

DAFTAR PUSTAKA

[1] N. Dwitri, J. A. Tampubolon, S. Prayoga, F. Ilmi Zer, and D. Hartama, "Penerapan Algoritma K-Means Dalam Menentukan Tingkat Kepuasan Pembelajaran Online Pada Masa Pandemi Covid-19 di Indonesia," *Jti (Jurnal Teknol. Informasi)*, vol. 4, no. 1, pp. 101–105, 2020.

[2] R. A. Indraputra and R. Fitriana, "K-Means Clustering Data COVID-19," vol. 10, no. 3, pp. 275–282, 2020.

[3] A. Susilo *et al.*, "Coronavirus Disease 2019: Tinjauan Literatur Terkini," *J. Penyakit Dalam Indones.*, vol. 7, no. 1, p. 45, 2020, doi: 10.7454/jpdi.v7i1.415.

[4] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. Ilmi R.H.Zer, and D. Hartama, "Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia," *Jti (Jurnal Teknol. Informasi)*, vol. 4, no. 1, pp. 166–173, 2020.

[5] A. P. Windarto, U. Indriani, M. R. Raharjo, and L. S. Dewi, "Bagian 1: Kombinasi Metode Klastering 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.

[6] I. G. I. Sudipa, I Nyoman Alit Arsana, and Made Leo Radhitya, "Penentuan Tingkat Pemahaman Mahasiswa Terhadap Social Distancing Menggunakan Algoritma C4.5," *SINTECH (Science Inf. Technol. J.*, vol. 3, no. 1, pp. 1–7, 2020, doi: 10.31598/sintechjournal.v3i1.562.

[7] A. F. Muhammad, "Klasterisasi Proses Seleksi Pemain Menggunakan Algoritma K-Means (Study Kasus : Tim Hockey Kabupaten Kendal)," *Jur. Tek. Inform. FIK UDINUS*, pp. 1–5, 2015.

[8] S. Turnip and P. Siltionga, "Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5," *J. Tek. Inform. Unika St. Thomas*, vol. 03, no. 479, pp. 3–7, 2018.