

Analisis *Clustering* Provinsi Indonesia Berdasarkan Persebaran Virus Corona (Covid-19) Menggunakan Algoritma *K-Means*

Vina Fitriyani¹

Universitas Negeri Semarang¹
vinafitriyani396@gmail.com

Abstrak—Virus Corona (Covid-19) telah menyebar ke seluruh benua dan terjadi di setiap negara di dunia. Virus Corona menjadi kasus yang banyak dibicarakan hampir di semua negara khususnya di Indonesia. Persebaran Virus Corona (Covid-19) tiap provinsi di Indonesia berbeda-beda. Salah satu upaya yang dapat dilakukan oleh para peneliti yaitu perlu melakukan suatu penelitian terkait pengelompokan provinsi berdasarkan persebaran Virus Corona (Covid-19) menggunakan algoritma *K-Means clustering* yang menjadi hal penting agar memudahkan dalam membedakan kondisi tiap provinsi sehingga dapat melahirkan suatu kebijakan yang lebih tepat. Data dari penelitian ini bersumber dari salah satu website yakni <http://kawalcovid19.id>. Pada penelitian ini digunakan sebanyak 34 provinsi di Indonesia. Atribut yang digunakan ada 10, yaitu nama provinsi, Orang Tanpa Gejala (OTG) total, Orang Dalam Pemantauan (ODP) total, Pasien Dalam Pengawasan (PDP), *suspect* total, *probable* total, kasus dirawat, kasus sembuh, dan kasus meninggal. Data yang telah dikumpulkan kemudian dilakukan *clustering* dengan menentukan *k cluster* menggunakan algoritma *hierarchical clustering* dan didapatkan sebanyak 6 *cluster*, yaitu *cluster 0*, *cluster 1*, *cluster 2*, *cluster 3*, *cluster 4*, dan *cluster 5*. Masing-masing *cluster* dikategorikan ke dalam *cluster* sangat rendah, cukup rendah, rendah, tinggi, cukup tinggi, dan sangat tinggi. Dalam implementasinya menggunakan *software R studio* untuk menentukan *k cluster* dan *Rapidminer studio version 9.6* untuk menjalankan algoritma *K-Means clustering*. Sehingga diperoleh hasil untuk persebaran Virus Corona di Indonesia berdasarkan provinsi *cluster 0* sebanyak 12 provinsi, *cluster 1* sebanyak 1 provinsi, *cluster 2* sebanyak 1 provinsi, *cluster 3* sebanyak 1 provinsi, *cluster 4* sebanyak 17 provinsi, dan *cluster 5* sebanyak 2 provinsi.

Kata kunci: Covid-19, Algoritma clustering, K-Means

I. PENDAHULUAN

Virus Corona (Covid-19) telah menyebar ke seluruh benua dan terjadi di setiap negara di dunia. Cepatnya penyebaran dan banyaknya pasien adalah kesulitan utama dalam mitigasi bencana ini. Penyakit ini merupakan penyakit yang dapat menular yang dibawa oleh berbagai macam mikroba, seperti bakteri, parasit, jamur, dan virus. Virus Corona (Covid-19) menjadi kasus yang banyak dibicarakan hampir di semua negara khususnya di Indonesia. Virus Corona (Covid-19) ini pertama kali terdeteksi di Wuhan, salah satu kota yang berada di Tiongkok, provinsi Hubei, Cina. Pandemi Covid-19 yang sedang berlangsung dari akhir tahun 2019 yakni 1 Desember 2019 hingga sekarang ini disebabkan oleh sindrom pernafasan akut yang parah Virus Corona 2 (SARS-CoV-2) yang bersumber dari kelelawar atau mungkin trenggiling. Pandemi Covid-19 mulai menyebar ke Indonesia pada 2 Maret 2020 hingga sekarang telah menyebar ke semua provinsi.

Terlepas dari berbagai opsi kebijakan yang ditempuh pemerintah terkait pencegahan persebaran Virus Corona diantaranya dengan menjaga jarak fisik, menggunakan masker, dan senantiasa menjaga kebersihan dengan mencuci tangan menggunakan sabun dan air mengalir. Namun, jumlah korban akibat Virus Corona semakin bertambah dengan munculnya *cluster* baru. Persebaran Virus Corona (Covid-19) tiap provinsi di Indonesia berbeda-beda. Hal ini dapat dilihat dari banyaknya jumlah Orang Tanpa Gejala (OTG) total, Orang Dalam Pemantauan (ODP) total, Pasien Dalam Pengawasan (PDP), *suspect* total, *probable* total, kasus dirawat, kasus sembuh, dan kasus meninggal dari tiap provinsi. Salah satu upaya yang dapat dilakukan oleh para peneliti yaitu perlu melakukan suatu penelitian terkait pengelompokan terhadap kasus Virus Corona

(Covid-19) dengan harapan dapat membantu pengambilan keputusan yang tepat dalam mengurangi persebaran Virus Corona karena wabah Covid-19 ini merupakan persoalan serius di dunia terutama Indonesia yang dapat merenggut banyak nyawa. Selain itu pengelompokan ini dilakukan untuk sarana proyeksi dan optimasi pemerintah dalam mengambil kebijakan dan mengevaluasi kebijakan yang ada serta menjadi hal penting agar memudahkan dalam membedakan kondisi tiap provinsi sehingga dapat melahirkan suatu kebijakan yang lebih tepat.

Dalam dunia komputasi, semakin maraknya teknologi dan informasi di masa sekarang ini mendorong adanya inovasi teknologi informasi dan media penyimpanan data yang telah memungkinkan manusia dalam melakukan pengumpulan dan menyiapkan data dari berbagai sumber yang tersedia. Data dan informasi menjadi bagian yang penting dalam mempertimbangkan pengambilan keputusan terkait penanganan suatu kasus. Data mining memungkinkan orang dalam menganalisa data untuk digali informasi tersembunyi dan dapat membantu dalam membuat keputusan berdasarkan informasi yang telah diperoleh. Salah satu metode yang diterapkan dalam data mining yaitu *clustering* (pengelompokan). *Clustering* akan melakukan pengelompokan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada[2]. Dalam masa pandemi ini *clustering* dapat digunakan untuk mengelompokan provinsi Indonesia berdasarkan persebaran Virus Corona (Covid-19) dengan dilengkapi suatu atribut diantaranya nama provinsi, Orang Tanpa Gejala (OTG) total, Orang Dalam Pemantauan (ODP) total, Pasien Dalam Pengawasan (PDP), *suspect* total, *probable* total, kasus dirawat, kasus sembuh, dan kasus meninggal.

Pada penelitian ini akan dilakukan suatu pengelompokan dengan menggunakan algoritma *K-Means clustering* yang digabungkan dengan algoritma *hierarchical clustering* untuk menentukan nilai k. Nilai k disini merupakan banyaknya *cluster* yang akan dibentuk. Kemudian tetapkan nilai-nilai k secara random, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid*, *mean* atau “*means*”[1]. Algoritma *K-Means* dipilih karena kemudahannya dan kemampuannya untuk mengklaster data yang besar dan data *outlier* dengan sangat cepat. Kelemahan *K-Means clustering* terdapat pada penentuan pusat awal *cluster* dilakukan secara random. Sehingga akan lebih optimal dengan menggabungkan algoritma *K-Means clustering* dengan *hierarchical clustering*. Dalam pengolahan data akan digunakan *software R studio* dan *Rapidminer studio version 9.6*. Metode pengelompokan *K-Means* digunakan untuk mengelompokkan data-data yang memiliki ciri yang sama dan mengelompokkannya ke dalam sebuah *cluster*[3]. Oleh karena itu dengan menggunakan salah satu metode *clustering* penelitian ini dapat mengelompokkan provinsi Indonesia berdasarkan persebaran Virus Corona (Covid-19) dengan menggunakan algoritma *K-Means*.

II. METODE PENELITIAN

Metode penelitian digunakan sebagai acuan dalam melakukan penelitian. Pada penelitian ini tahapan yang digunakan yaitu mempelajari literatur, mengumpulkan data, *pre-processing* data, implementasi *software*, dan menganalisa hasil.

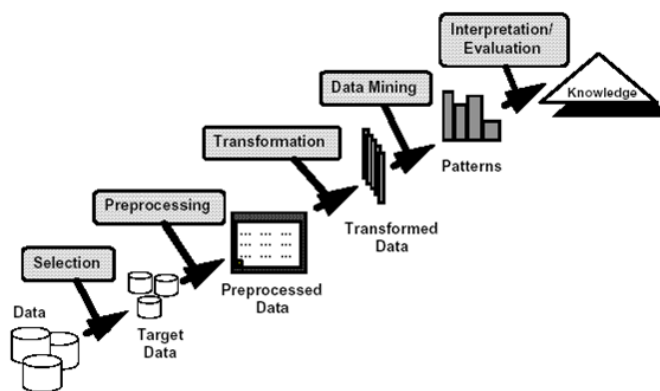
A. Dataset

Dataset yang digunakan dalam penelitian ini berupa data sekunder yaitu persebaran Virus Corona (Covid-19) di Indonesia yang disediakan berdasarkan provinsi yang diambil pada tanggal 14 Maret 2021. Data yang digunakan bertipe data kuantitatif yang diperoleh dari salah satu website yakni <https://kawalcovid19.id> yang dikelola oleh Tim Relawan Kawal Covid-19. Kemudian data tersebut diseleksi dan akan digunakan untuk diolah dalam mengelompokkan banyaknya provinsi menurut persebaran Virus Corona (Covid-19) di Indonesia. Adapun atribut yang digunakan pada penelitian ini yaitu nama provinsi, Orang Tanpa Gejala (OTG) total, Orang Dalam Pemantauan (ODP) total, Pasien Dalam Pengawasan (PDP), *suspect* total, *probable* total, kasus dirawat, kasus sembuh, dan kasus meninggal.

B. Data Mining

Data mining merupakan proses untuk menemukan karakteristik yang penting dalam data yang sedang digunakan dan menemukan model untuk menggambarkan kelas atau suatu konsep data[5]. Proses-proses dalam data mining digunakan untuk mencari nilai tambah dari suatu kumpulan data berupa informasi yang sebelumnya tidak pernah diketahui secara manual. Ada beberapa ilmu yang mendukung teknik data mining diantaranya adalah data analisis, *signal processing*, *neural network*, dan pengolahan pola[4]. Secara garis besar data mining dikelompokkan menjadi 2 kategori utama yaitu *descriptive* mining dan *predictive*[1]. Teknik data mining yang merupakan *descriptive* mining antara lain *clustering*, *association*, dan *sequential mining* sedangkan salah satu teknik data mining yang termasuk ke dalam *predictive* adalah klasifikasi.

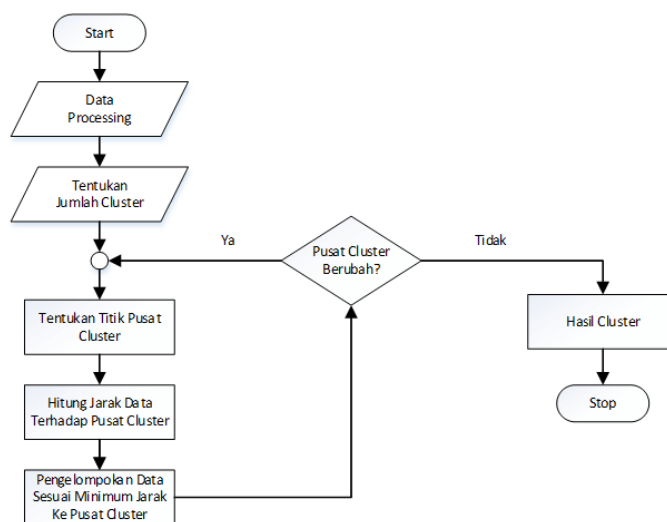
Adapun istilah lain untuk data mining adalah *Knowledge Discovery in Database* (KDD). Data mining termasuk bagian dalam proses KDD seperti yang ditunjukkan pada Gambar 1.



GAMBAR 1. TAHAPAN KNOWLEDGE DISCOVERY IN DATABASE (KDD)

C. Clustering

Clustering dapat diartikan proses untuk mengorganisasikan sekelompok data ke dalam berbagai kelompok-kelompok sedemikian rupa sehingga objek-objek yang serupa akan menjadi satu *cluster* sedangkan objek-objek yang tidak serupa menjadi anggota *cluster* yang lain[5]. Potensi *clustering* adalah dapat digunakan untuk mengetahui struktur dalam data yang dapat dipakai lebih lanjut dalam berbagai aplikasi secara luas seperti klasifikasi pengolahan gambar, dan pengolahan pola[4]. Sedangkan peneliti lain yakni (Rahmat *et al.* (2017)) mengatakan, *clustering* adalah membagi objek data (bentuk, entitas, contoh, ketaatan, unit) ke dalam beberapa jumlah kelompok (grup, bagian atau kategori). *Flowchart* dari proses *clustering* ditunjukkan pada Gambar 2.



GAMBAR 2. FLOWCHART CLUSTERING

D. Algoritma K-Means

Algoritma *K-Means Clustering* merupakan salah satu metode penganalisaan data mining yang bersifat *unsupervised learning* dengan sistem partisi. Algoritma *K-Means* memiliki tujuan dalam pengelompokan data dengan meminimalkan data antar *cluster* dan memaksimalkan data dalam satu *cluster*. Prinsip utama dari teknik ini adalah menyusun k buah partisi/pusat massa (*centroid*)/rata-rata (*mean*) dari sekumpulan data[6]. Variabel k sendiri merupakan jumlah *cluster* yang diinginkan. Algoritma *K-Means* menerima penginputan berupa data tanpa label/kelas. Ukuran kemiripan yang digunakan dalam *clustering* adalah fungsi jarak. Sehingga untuk memaksimalkan kemiripan data diambil berdasarkan jarak terpendek antar data terhadap titik *centroid*.

Tahap awal dalam melakukan proses pengklasteran data menggunakan algoritma *K-Means* yaitu dengan pembentukan titik awal *centroid* (c_j). Pada umumnya titik *centroid* (c_j) dibangkitkan secara acak dari

sekumpulan populasi data. Selanjutnya jumlah *centroid* c_j yang dibangkitkan akan sesuai dengan jumlah *cluster* yang ditentukan diawal. Setelah k *centroid* terbentuk kemudian dihitung jarak tiap data x_i dengan *centroid* ke- j sampai k dinotasikan dengan $d(x_i, c_j)$. Adapun ukuran kedekatan data dihitung menggunakan jarak *euclidean*. Perhitungan jarak *euclidean* seperti pada Persamaan 1.

$$d(X_i, C_j) = \sqrt{\sum_{i=1}^n (X_i - C_j)^2} \quad (1)$$

Keterangan:

- $d(X_i, C_j)$: jarak antara data ke i dan data ke j .
- n : dimensi data
- X_i : koordinat data ke i
- C_j : koordinat data ke j

E. Rapidminer

Rapidminer merupakan *software*/perangkat lunak untuk pengolah data yang mempunyai sifat terbuka (*open source*). *Rapidminer* merupakan sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. Hal ini digunakan untuk bisnis dan komersial, penelitian, pendidikan, pelatihan, *rapid prototyping*, dan pengembangan aplikasi serta mendukung semua langkah dalam proses pembelajaran mesin termasuk persiapan data, hasil visualisasi, validasi model, dan optimasi. *Rapidminer* sebelumnya bernama YALE (*Yet Another Learning Environment*). Dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mirswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. Dengan menggunakan *Rapidminer*, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan[5]. *Rapidminer* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang baik[6].

III. HASIL DAN PEMBAHASAN

A. Tahap Pengumpulan Data

Untuk melakukan penelitian ini, langkah pertama yang dilakukan adalah mengumpulkan data yang akan digunakan. Dalam penerapan algoritma *K-Means clustering* untuk mengelompokkan banyaknya provinsi menurut persebaran Virus Corona (Covid-19) di Indonesia diperlukan data terkait mengenai hal itu. Dataset persebaran Virus Corona (Covid-19) di Indonesia berdasarkan provinsi terdiri dari beberapa atribut yang dapat dilihat pada Tabel 1.

TABEL 1. STRUKTUR DATASET PERSEBARAN VIRUS CORONA (COVID-19) DI INDONESIA BERDASARKAN PROVINSI

Nama field	Tipe	Keterangan
Provinsi	String	Nama provinsi
OTG total	Numerik	Jumlah orang tanpa gejala
ODP total	Numerik	Jumlah orang dalam pemantauan
PDP	Numerik	Jumlah pasien dalam pengawasan
<i>Suspect</i> total	Numerik	Jumlah orang yang diduga kuat terjangkit Covid-19
<i>Probable</i> total	Numerik	Jumlah orang yang yang diyakini sebagai <i>suspect</i> dengan ISPA berat atau gagal napas
Kasus dirawat	Numerik	Jumlah pasien yang dirawat
Kasus sembuh	Numerik	Jumlah pasien yang sembuh

Kasus meninggal	Numerik	Jumlah pasien yang meninggal
-----------------	---------	------------------------------

B. Tahap Pre-Processing Data

Sebelum melakukan penambahan data menggunakan *software Rapidminer studio* terlebih dahulu dilakukan *pre-processing* data yaitu data perlu dibersihkan sebelum diproses, hal ini terjadi karena biasanya data yang akan digunakan belum baik. Teknik atau metode yang digunakan dalam melakukan *pre-processing* data diantaranya adalah data *cleaning*, data *integration*, data *transformation*, dan data *reduction*.

Pada tahap data *cleaning* data yang tidak relevan, *missing value*, dan duplikat harus dibersihkan. Hal ini dikarenakan data yang relevan, tidak *missing value*, dan tidak duplikat merupakan syarat awal dalam melakukan proses data mining. Suatu data dikatakan *missing value* jika terdapat atribut dalam dataset yang tidak berisi nilai atau kosong, sedangkan suatu data dikatakan duplikat jika dalam satu dataset lebih dari satu *record* yang berisi nilai yang sama. Pada dataset yang digunakan terdapat *missing value* yang ditunjukkan pada Tabel 2 diantaranya atribut OTG total terdapat *missing value* sebanyak 18 data, atribut ODP total terdapat *missing value* sebanyak 22 data, atribut PDP terdapat *missing value* sebanyak 24 data, atribut *Suspect* total terdapat *missing value* sebanyak 12 data, dan atribut *Probable* total terdapat *missing value* sebanyak 14 data, serta untuk atribut kasus dirawat, kasus sembuh dan kasus meninggal tidak terdapat *missing value* yang dapat dilihat pada Tabel 2.

TABEL 2. DATA SEBELUM DILAKUKAN *REPLACE MISSING VALUE*

Provinsi	OTG Total	ODP Total	PDP	Suspect Total	Probable Total	Kasus Dirawat	Kasus Sembuh	Kasus Meninggal
Aceh				7087	672	1385	7911	385
Bali						1799	34084	1017
Banten	46125			28108	54	3076	35297	1064
Bangka Belitung	19898			6026	32	564	7285	124
Bengkulu		7464				278	4697	147
DI Yogyakarta		1507	212	35098	9845	3561	25293	883
DKI Jakarta	972395			753779	7698	3357	345202	5991
Jambi				3283		985	4661	88
Jawa Barat	323706			143878	4107	16464	209026	4982
Jawa Tengah	134739			74099	3244	7219	181206	11542
Jawa Timur	777927			75846	979	1147	122952	9451
Kalimantan Barat	6	24040	1779			610	4505	33
Kalimantan Timur				180539	30	5386	52855	1404
Kalimantan Tengah				11105	106	1371	13298	390
Kalimantan Selatan		5091	987			2297	21064	769
Kalimantan Utara	16026			10012	75	1307	8936	160
Kepulauan Riau						141	8483	222
Nusa Tenggara Barat	66751			17458	93	964	8568	413
.
.
Sulawesi Barat	2599	1978	152			54	4534	110
Nusa Tenggara Timur	13142			3331	127	2145	8267	299
Gorontalo	8976			4602	3	159	4653	142

Setelah dilakukan *replace missing value* dengan metode menghitung nilai pengganti (*imputation*) yaitu melengkapi data yang kosong atau *missing value* dengan nilai rata-rata menggunakan program *Rapidminer* dan diperoleh seperti pada Tabel 3.

TABEL 3. DATA SETELAH DILAKUKAN *REPLACE MISSING VALUE*

Provinsi	OTG Total	ODP Total	PDP	Suspect Total	Probable Total	Kasus Dirawat	Kasus Sembuh	Kasus Meninggal
Aceh	152772	14406	1075	7087	672	1385	7911	385
Bali	152772	14406	1075	67877	1413	1799	34084	1017
Banten	46125	14406	1075	28108	54	3076	35297	1064
Bangka Belitung	19898	14406	1075	6026	32	564	7285	124
Bengkulu	152772	7464	1075	67877	1413	278	4697	147
DI Yogyakarta	152772	1507	212	35098	9845	3561	25293	883
DKI Jakarta	972395	14406	1075	753779	7698	3357	345202	5991
Jambi	152772	14406	1075	3283	1413	985	4661	88
Jawa Barat	323706	14406	1075	143878	4107	16464	209026	4982
Jawa Tengah	134739	14406	1075	74099	3244	7219	181206	11542
Jawa Timur	777927	14406	1075	75846	979	1147	122952	9451
Kalimantan Barat	6	24040	1779	67877	1413	610	4505	33
Kalimantan Timur	152772	14406	1075	180539	30	5386	52855	1404
Kalimantan Tengah	152772	14406	1075	11105	106	1371	13298	390
Kalimantan Selatan	152772	5091	987	67877	1413	2297	21064	769
Kalimantan Utara	16026	14406	1075	10012	75	1307	8936	160
Kepulauan Riau	152772	14406	1075	67877	1413	141	8483	222
Nusa Tenggara Barat	66751	14406	1075	17458	93	964	8568	413
.
.
.
Sulawesi Barat	2599	1978	152	67877	1413	54	4534	110
Nusa Tenggara Timur	13142	14406	1075	3331	127	2145	8267	299
Gorontalo	8976	14406	1075	4602	3	159	4653	142

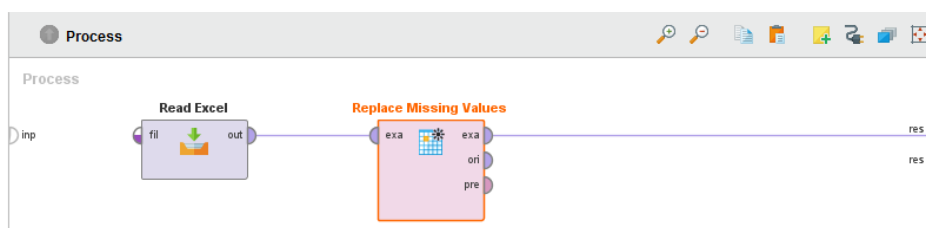
Selanjutnya untuk tahap data *integration* ini merupakan penggabungan data dari berbagai database yang berbeda. Pada tahapan ini tidak ada penggabungan data karena data yang diambil berasal dari satu database.

Tahap selanjutnya yaitu data *transformation* yang dilakukan untuk mengubah tipe data pada atribut dataset. Pada penelitian ini dataset yang digunakan tidak terdapat tipe data yang perlu diubah.

Selanjutnya data *reduction* dilakukan untuk menguraikan data ke dalam bentuk yang lebih kecil tetapi tetap menghasilkan analisis yang sama[3]. Pada dataset yang digunakan memiliki 37 atribut. Namun setelah dilakukan data *reduction* maka yang akan digunakan hanya 10 atribut yakni nama provinsi, Orang Tanpa Gejala (OTG) total, Orang Dalam Pemantauan (ODP) total, Pasien Dalam Pengawasan (PDP), *suspect* total, *probable* total, kasus dirawat, kasus sembuh, dan kasus meninggal.

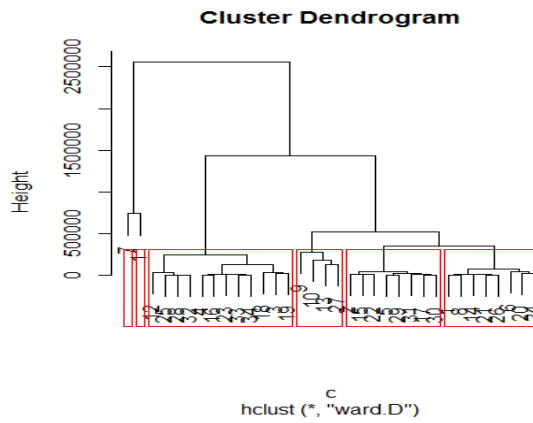
C. Tahap Penambangan Data

Pada tahap ini dilakukan tiga kali proses data mining yang pertama proses *replace missing value*, kedua menentukan nilai k dengan algoritma *hierarchical clustering*, dan ketiga proses *clustering* (pengelompokan) provinsi di Indonesia berdasarkan persebaran Virus Corona (Covid-19) menggunakan algoritma *K-Means clustering*. Proses *replace missing value* ditunjukkan pada konfigurasi operator seperti pada Gambar 3.



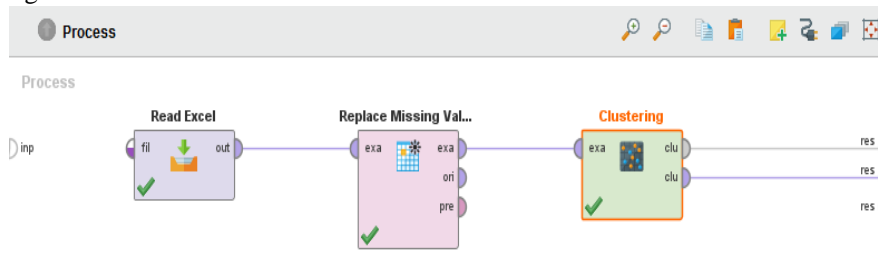
GAMBAR 3. PROSES REPLACE MISSING VALUE

Dengan konfigurasi operator, sebelum dataset persebaran Virus Corona (Covid-19) di Indonesia berdasarkan provinsi dilakukan proses pengelompokan maka terlebih dahulu dilakukan *replace missing value* untuk melengkapi data yang kosong dengan memasukkan nilai parameter *default* berupa *average* artinya data yang kosong akan dilengkapi dengan nilai *average*. Sebelum melanjutkan ke tahap *clustering* terlebih dahulu menentukan nilai k menggunakan algoritma *hierarchical clustering* metode *ward* dengan bantuan software *R studio* dan diperoleh nilai k sebanyak 6 *cluster* yang dapat dilihat pada Gambar 4.



GAMBAR 4. CLUSTER DENDROGRAM DENGAN ALGORITMA HIERARCHICAL CLUSTERING

Setelah didapat nilai k, tahap selanjutnya yaitu melakukan *clustering* dengan memasukkan dataset terlebih dahulu ke dalam operator *read excel* yang telah diletakkan pada bidang proses. Selanjutnya letakkan operator *K-Means clustering* ke dalam bidang proses. Pada parameter *K-Means clustering* ganti nilai k yang semula bernilai 5 diubah menjadi 6 karena pada penelitian ini akan dikelompokkan menjadi 6 *cluster*. Setelah melakukan jumlah *cluster* maka operator *read excel* dihubungkan dengan operator *K-Means clustering* untuk mengetahui *output* ke arah *result* seperti pada Gambar 5. Selanjutnya jalankan program untuk mendapatkan hasil dari pengolahan dataset.

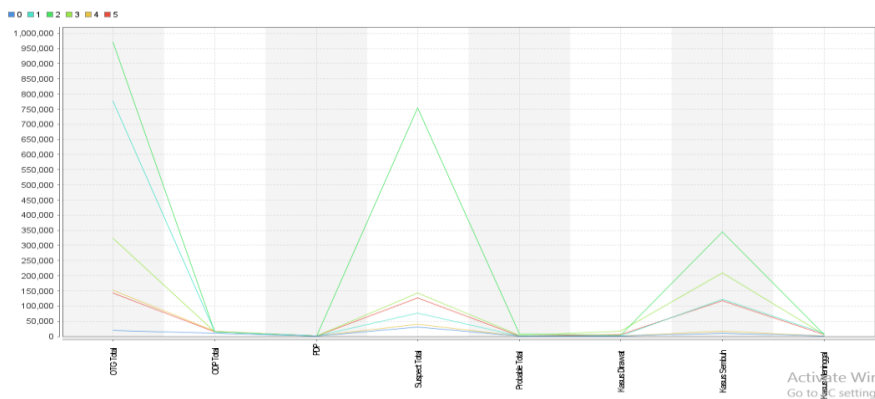


GAMBAR 5. PROSES CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS

Adapun hubungan setiap titik *centroid* dalam menentukan banyaknya jumlah data yang ada pada setiap *cluster* dapat dilihat pada Gambar 6 dan Gambar 7.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
OTG Total	19632.333	777927	972395	323706	152772	143755.500
ODP Total	11216.333	14406	14406	14406	16657.294	14406
PDP	863.667	1075	1075	1075	1223.941	1075
Suspect Total	31974.667	75846	753779	143878	40939.706	127319
Probable Total	535.667	979	7698	4107	1502.765	1637
Kasus Dirawat	969.167	1147	3357	16464	1750.353	6302.500
Kasus Sembuh	9891.750	122952	345202	209026	17836.706	117030.500
Kasus Meninggal	311.250	9451	5991	4982	524.118	6473

GAMBAR 6. TABEL HUBUNGAN CENTROID DARI TIAP CLUSTER



. GAMBAR 7. GRAFIK HUBUNGAN CENTROID DARI TIAP CLUSTER

Analisis pengelompokan penyebaran Virus Corona (Covid-19) di Indonesia berdasarkan provinsi menggunakan *Rapidminer studio* dapat dilakukan dengan mudah. Kita dapat mengetahui provinsi mana saja yang masuk ke dalam *cluster 0* atau *cluster* sangat rendah, *cluster 1* atau *cluster* cukup rendah, *cluster 2* atau *cluster* rendah, *cluster 3* atau *cluster* tinggi, *cluster 4* atau *cluster* cukup tinggi, maupun *cluster 5* atau *cluster* sangat tinggi dapat dilihat melalui *output* pada dataset yang telah diolah. Hasil *clustering* data menunjukkan 34 data yang dikelompokkan menjadi 6 *cluster* diantaranya *cluster 0* atau *cluster* sangat rendah berjumlah 12 provinsi yaitu provinsi Banten, Bangka Belitung, Kalimantan Barat, Kalimantan Utara, Nusa Tenggara Barat, Sumatra Selatan, Sulawesi Tenggara, Sulawesi Tengah, Maluku Utara, Sulawesi Barat, Nusa Tenggara Timur, dan Gorontalo, *cluster 1* atau *cluster* cukup rendah berjumlah 1 provinsi yakni provinsi Jawa Timur, *cluster 2* atau *cluster* rendah berjumlah 1 provinsi yakni provinsi DKI Jakarta, *cluster 3* atau *cluster* tinggi berjumlah 1 provinsi yakni provinsi Jawa Barat, *cluster 4* atau *cluster* cukup tinggi berjumlah 17 provinsi yakni Aceh, Bali, Bengkulu, DI Yogyakarta, Jambi, Kalimantan Tengah, Kalimantan Selatan, Kepulauan Riau, Sumatra Barat, Sulawesi Utara Sumtra Utara, Sulawesi Selatan, Lampung, Riau, Maluku, Papua Barat, dan Papua, serta untuk *cluster 5* atau *cluster* sangat tinggi berjumlah 2 provinsi yakni provinsi Jawa Tengah dan Kalimantan Timur.

IV. SIMPULAN DAN SARAN

Untuk melakukan pengelompokan provinsi di Indonesia menurut persebaran virus Corona (Covid-19) dapat menerapkan metode *K-Means Clustering*. Data diolah untuk ditentukan nilai *centroid* dari masing-masing atribut dalam 6 *cluster* yaitu *cluster 0* atau *cluster* rendah, *cluster 1* atau *cluster* sedang, dan *cluster 2* atau *cluster* tinggi. Diperoleh 12 provinsi yang masuk ke dalam *Cluster 0* atau *cluster* sangat rendah yaitu provinsi Banten, Bangka Belitung, Kalimantan Barat, Kalimantan Utara, Nusa Tenggara Barat, Sumatra Selatan, Sulawesi Tenggara, Sulawesi Tengah, Maluku Utara, Sulawesi Barat, Nusa Tenggara Timur, dan Gorontalo, 1 provinsi yang masuk pada *cluster 1* atau *cluster* cukup rendah yakni provinsi Jawa Timur, 1 provinsi yang masuk ke dalam *cluster 2* atau *cluster* rendah yakni provinsi DKI Jakarta, 1 provinsi yang masuk ke dalam *cluster 3* atau *cluster* tinggi yakni provinsi Jawa Barat, 17 provinsi yang masuk ke dalam *cluster 4* atau *cluster* cukup tinggi yakni Aceh, Bali, Bengkulu, DI Yogyakarta, Jambi, Kalimantan Tengah, Kalimantan Selatan, Kepulauan Riau, Sumatra Barat, Sulawesi Utara Sumtra Utara, Sulawesi Selatan, Lampung, Riau, Maluku, Papua Barat, dan Papua, serta 2 provinsi yang masuk ke dalam *cluster 5* atau *cluster* sangat tinggi yakni provinsi Jawa Tengah dan Kalimantan Timur.

Peneliti menyadari adanya kekurangan dari penelitian ini karena keterbatasan peneliti baik dalam hal waktu maupun pengetahuan. Peneliti memberikan beberapa saran untuk dijadikan masukan bagi pemerintah khususnya provinsi, agar provinsi yang masuk ke dalam kategori *cluster* tinggi, cukup tinggi, dan sangat tinggi mendapat perhatian lebih dan meningkatnya kebijakan dalam usaha pencegahan penyebaran Covid-19 di Indonesia. Selain itu disarankan untuk menggunakan algoritma yang berbeda dalam pengelompokan provinsi Indonesia berdasarkan persebaran Virus Corona (Covid-19) dan terdapat beberapa metode *clustering* lainnya yang perlu diterapkan untuk dijadikan sebagai referensi guna mengembangkan metode *K-Means* serta akan lebih menarik apabila ditampilkan peta *cluster* yang menggambarkan pengelompokan provinsi Indonesia berdasarkan persebaran Virus Corona (Covid-19).

DAFTAR PUSTAKA

- [1] Darmi, Y., and Setiawan, A. "Penerapan metode *clustering K-Means* dalam pengelompokan penjualan produk," *Jurnal Media Infotama Universitas Muhammadiyah Bengkulu*, 12(2), 148–157, September 2016.
- [2] Dhuhita, W. M. P. "Clustering Menggunakan Metode *K-Means* Untuk," *Jurnal Informatika*, 15(2), 160–174, Desember 2015.
- [3] Rahmat C.T.I., B., Agidama Gafar, A., Fajriani, N., Ramdani, U., Rihin Uyun, F., Purnamasari P., Y., dkk. "Implementasi *K-Means clustering* pada *Rapidminer* untuk analisis daerah rawan kecelakaan," *Seminar Nasional Riset Kuantitatif Terapan 2017*, 58–60, April 2017.
- [4] Sari, R. W., Wanto, A., and Windarto, A. P. "Implementasi *Rapidminer* Dengan Metode *K-Means* (Study Kasus: Imunisasi Campak Pada Balita Berdasarkan Provinsi)," *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 2(1), 224–230, Oktober 2018
- [5] Tendean, T., Purba, W., and Kom, M. "Analisis *Cluster* Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma *K-Means*," 1(2), 5–11, Maret 2020.
- [6] Sibuea, F. L., Sapta, A., Informasi, S., and Royal, S. "PEMETAAN SISWA BERPRESTASI MENGGUNAKAN METODE *K-MEANS CLUSTERING*," 1(1), 85-92, Desember 2017.