ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 1, Tahun 2022, Halaman 56 – 66
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

KLASTERISASI PROVINSI DI INDONESIA BERDASARKAN FAKTOR PENYEBARAN COVID-19 MENGGUNAKAN *MODEL-BASED CLUSTERING t*-MULTIVARIAT

Nor Hamidah[1], Rukun Santoso[2*], Agus Rusgiyono[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***email :** rukunsantoso25@gmail.com

ABSTRACT

The spread of Covid-19 had a significant impact in all sectors. Enforcement policies from the government that are appropriate with the conditions for the spread of the virus that are needed to prevent a bigger impact. Clusteritation by province based on data on the spread of Covid-19 is important for the government to set appropriate policies in order to prevent the spread of Covid-19. The data used include data on population density, testing rate, proportion of population 50 years and over, and proportion of population diligently hand-washing in each province. The data factors for the spread of Covid-19 tend to overlap and there are outliers in the data which causes the data not normally distributed. In this study, *Model-Based Clustering t*-multivariate was used for data clustering. The results show that using Integrated Completed Likelihood, two groups of optimal cluster were obtained. The second cluster has a higher risk of spreading Covid-19 than the first cluster.

Keywords : Covid-19, Clustering, *Model-Based Clustering t*-Multivariat

1. PENDAHULUAN

Penyebaran Covid-19 memberikan dampak yang signifikan hampir di semua sektor. Di sektor kesehatan, pandemi Covid-19 mempengaruhi kondisi fisik dan mental seseorang. Depresi, kecemasan dan kekhawatiran yang berlebihan, serta kelelahan banyak dialami masyarakat khususnya tenaga medis (Rosyanti dan Hadi, 2020). Di sektor ekonomi, pandemi Covid-19 membuat banyak karyawan terkena Pemutusan Hubungan Kerja (PHK) dan beberapa perusahaan terancam bangkrut (Yamali dan Putri, 2020). Pandemi Covid-19 juga mempengaruhi kegiatan ekspor-impor di Indonesia. Pada bulan Januari 2020, ekspor migas mengalami penurunan sebesar 9,15% dan impor migas mengalami penurunan sebesar 3,08% dibandingkan Desember 2019. Di sektor pariwisata, pembatasan aktivitas warga di berbagai negara menyebabkan penurunan jumlah wisatawan. Di Indonesia, salah satu provinsi yang terdampak paling besar adalah provinsi Bali. Secara keseluruhan, penurunan wisatawan di Bali karena pandemi Covid-19 mencapai angka 50% (Budiyanti, 2020).

Pemberlakuan kebijakan dari pemerintah yang sesuai dengan kondisi faktor penyebaran virus diperlukan untuk mencegah terjadinya dampak yang lebih besar. Pengelompokan wilayah berdasarkan data faktor penyebaran Covid-19 penting dilakukan oleh pemerintah untuk menetapkan kebijakan yang sesuai dalam rangka pencegahan penyebaran Covid-19. Pengelompokan wilayah dapat dilakukan dengan analisis klaster.

Analisis klaster adalah metode statistik yang digunakan untuk membentuk kelompok-kelompok (klaster) dari suatu data multivariat. Data faktor penyebaran Covid-19 memiliki karakteristik cenderung *overlap* (tumpang tindih) dan terdapat *outlier* (pencilan) pada data yang menyebabkan data tidak berdistribusi normal. Data yang *overlap* sulit untuk dikelompokkan menggunakan metode klastering berbasis jarak. Untuk mengatasi hal ini, dapat digunakan *Model-Based Clustering t-multivariat.* Distribusi *t* dipilih karena lebih *robust* terhadap *outlier*. Kriteria *Integrated Completed Likelihood* digunakan untuk menentukan jumlah klaster optimal.

2. TINJAUAN PUSTAKA

2.1. Covid-19

Covid-19 adalah penyakit menular yang disebabkan oleh jenis coronavirus yang baru ditemukan. Virus ini diduga muncul pertama kali di Wuhan, China. Laju kematian akibat Covid-19 memang lebih rendah dibandingkan SARS-CoV dan MERS-CoV. Namun, tingkat penyebarannya lebih cepat dibanding keduanya. Tingkat penyebaran virus yang tinggi disebabkan oleh cara penularannya. Faktor-faktor lain yang memiliki pengaruh terhadap tingginya penyebaran Covid-19 adalah:

1. Kepadatan penduduk

Penelitian oleh Roy dan Ghosh (2020) menyimpulkan bahwa kepadatan penduduk berpengaruh terhadap penyebaran virus secara signifikan. Semakin padat jumlah penduduk dalam suatu wilayah, peluang virus tersebar akan semakin tinggi. Kepadatan penduduk didefinisikan sebagai jumlah penduduk yang tinggal di suatu wilayah per kilometer persegi.

2. Testing Ratio

Peningkatan kasus positif Covid-19 berhubungan dengan jumlah pengujian yang dilakukan. Semakin banyak jumlah pengujian akan menyebabkan semakin banyak kasus yang terdeteksi. Apabila seseorang terkonfirmasi positif, maka akan segera dilakukan tindakan perawatan atau isolasi mandiri. Tindakan ini akan membuat peluang pasien yang positif Covid-19 untuk menularkan virus ke orang lain menjadi menurun (Velasco, Tseng dan Chang, 2021). Di Indonesia, beberapa tes yang umum digunakan untuk mendiagnostik virus corona yaitu Tes Cepat Molekuler (TCM), Real Time Polymerase Chain Reaction (RT PCR), dan Rapid Test.

3. Usia

Priyadarsini dan Suresh (2020) menyebutkan bahwa risiko penularan dan kematian akibat Covid-19 menjadi meningkat pada penduduk berusia lanjut. Imunitas yang lemah menyebabkan kelompok ini lebih rentan terpapar virus. Selain itu, mayoritas penduduk berusia lanjut sudah memiliki riwayat penyakit sebelumnya. Apabila terinfeksi, virus akan membuat penyakit yang diderita semakin parah.

4. Kebiasaan cuci tangan

Perilaku Hidup Bersih dan Sehat (PHBS) merupakan kunci penting dalam upaya pencegahan penyebaran Covid-19. Strategi ini dinilai cukup efektif dan mudah dilakukan oleh semua lapisan masyarakat. Di antara pola PHBS adalah mencuci tangan dengan bersih, konsumsi makanan sehat, olahraga, dan istirahat yang cukup (Karo, 2020).

2.2. Analisis Klaster

Analisis klaster merupakan salah satu metode dalam analisis multivariat yang bertujuan untuk melakukan pengelompokan objek-objek berdasarkan karakteristik yang dimilikinya. Tujuan dari analisis klaster adalah membentuk klaster yang homogen dari keseluruhan data yang heterogen. Klaster yang baik adalah klaster yang memiliki homogenitas yang tinggi antar anggota dalam satu klaster dan heterogenitas yang tinggi antar klaster.

Asumsi pada analisis klaster, seperti disebutkan oleh Hair *et al.* (2013) meliputi dua hal :

1. Sampel yang representatif

Sampel yang representatif diperlukan agar proses klastering memberikan hasil yang dapat mewakili populasinya. Untuk melakukan uji sampel yang representatif, dapat digunakan uji KMO *(Kaiser Meyer Olkin).*

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2}{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2 + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} \rho_{x_j x_k, x_l}^2} \quad (1)$$

dengan: p = banyaknya variabel

$r_{x_j x_k}$ = koefisien korelasi antara variabel ke- j dan ke- k

$\rho_{x_j x_k, x_l}$= koefisien korelasi parsial antara variabel ke-j dan k dengan variabel ke- l

Sampel dikatakan mewakili populasi jika nilai KMO lebih dari 0,5 (Kaiser, 1974).

2. Tidak terjadi multikolinearitas

Multikolinearitas adalah adanya hubungan linier yang sempurna atau pasti di antara beberapa atau semua variabel. Salah satu cara identifikasi adanya multikolinieritas adalah dengan menghitung nilai Faktor Inflasi Ragam (FIR, *Variance Inflation Factor* = VIF) dengan rumus sebagai berikut:

$$VIF = \frac{1}{1-{R_j}^2} \quad (2)$$

dengan ${R_j}^2$ adalah nilai koefisien determinasi variabel ke-j. ${R_j}^2$ diperoleh dengan melakukan regresi antara variabel ke-j dengan variabel lainnya. Multikolinieritas terjadi apabila nilai VIF > 10 (Hair *et al.*, 2013).

2.3. Identifikasi Pola Pengelompokkan Data

Identifikasi pola pengelompokkan data dilakukan untuk mengetahui kecenderungan jumlah klaster yang terbentuk. Pola pengelompokkan data dapat dilihat dari plot skor komponen utama. Jumlah komponen utama yang dipilih untuk diplotkan tergantung pada jumlah komponen utama yang mampu menjelaskan sebagian besar data.

Pola pengelompokkan data akan sulit diidentifikasi apabila data memiliki kondisi cenderung *overlap* (tumpang tindih). Suatu data memiliki kondisi *overlap* apabila terdapat lebih dari 1 objek pada area yang sama. Semakin banyak objek yang *overlap*, akan menyebabkan klaster yang dihasilkan juga *overlap* (Pardede dan Prasetyo, 2012).

Ukuran jarak tidak dapat dijadikan sebagai ukuran kesamaan antar objek dalam melakukan pengelompokkan pada klaster *overlap*. Penggunaan ukuran jarak tidak memiliki mekanisme untuk menentukan klaster yang tepat apabila objek dekat dengan lebih dari satu pusat klaster. Oleh karena itu, digunakan ukuran kesamaan lain yaitu berdasarkan distribusi datanya (Pardede dan Prasetyo, 2012).

2.4. *Model-Based Clustering t*-multivariat

Model-Based Clustering merupakan metode klastering yang menggunakan pendekatan model peluang untuk mengelompokkan data. Distribusi yang sering digunakan pada *Model-Based Clustering* adalah distribusi normal. Namun, tidak semua data memenuhi distribusi normal karena keberadaan *outlier*. Oleh karena itu, (Andrews dan McNicholas, 2012) mengembangkan model yang lebih *robust* pada data yang memiliki *outlier* dengan menggunakan distribusi *t*.

Pengujian normalitas multivariat dapat dilakukan dengan pendekatan grafik dan secara formal. Pengujian dengan pendekatan grafik menggunakan Q-Q plot *(Quantile-Quantile Plot).* Pengujian normalitas secara formal dilakukan dengan metode *Kolmogorov-Smirnov*.

- Hipotesis

H_0 : Data mengikuti distribusi normal multivariat

H_1: Data tidak mengikuti distribusi normal multivariat

- Statistik Uji

$$D = Sup|F(d^2{}_j) - F_0(d^2{}_j)| \quad (3)$$

dengan :

$F(d^2{}_j)$ = Proporsi jarak mahalanobis yang $\leq d^2{}_j$

$F_0(d^2{}_j)$= Fungsi peluang kumulatif dari distribusi *chi-square*

- Kriteria Uji

Tolak H_0 jika $D > W_{(1-\alpha)}$ dengan uji 2 sisi atau nilai *p-value*$< \alpha$, $W_{(1-\alpha)}$ merupakan kuantil $1 - \alpha$ pada tabel Kolmogorov-Smirnov.

Distribusi *t* merupakan distribusi peluang yang mirip dengan distribusi normal. Distribusi *t* memiliki bentuk kurva seperti distribusi normal, namun lebih gemuk di bagian ekornya. Bagian ekor yang gemuk menandakan terdapat objek dalam data yang cenderung menyimpang jauh dari rata-rata. Data yang menyimpang tersebut disebut dengan data *outlier* (pencilan). Jika suatu data memiliki banyak *outlier*, maka distribusi data akan menjadi lebih landai dan tidak berdistribusi normal. Untuk itu, dikembangkan distribusi *t* yang lebih *robust* dibandingkan dengan distribusi normal dalam mengatasi *outlier* (Agustini, 2017).

Pendeteksian *outlier* dilakukan apabila asumsi normalitas tidak terpenuhi. *Outlier* adalah data yang memiliki pola berbeda dengan sebagian besar pola data sehingga terletak jauh dari rata-ratanya. Pemeriksaan keberadaan *outlier* pada data multivariat dapat menggunakan jarak *Robust* (*Robust Distance).* Perhitungan RD menggunakan metode *Minimum Covariance Determinant* (MCD) untuk mengestimasi rata-rata dan matriks varian kovarian. *Outlier* dideteksi dan diklasifikasi secara visual dengan melihat *diagnostic plot. Diagnostic plot* adalah plot yang terbentuk antara jarak Mahalanobis dengan jarak *Robust.* Observasi pada *diagnostic plot* dibagi dalam 4 kuadran. Hanya pengamatan yang terletak di kuadran III yang termasuk dalam pengamatan biasa (bukan pengamatan *outlier*).

Parameter pada *Model-Based Clustering t-multivariat* dapat diestimasi menggunakan algoritma *Expectation-Conditional Maximization* (ECM). Langkah pertama pada algoritma ECM adalah langkah *Expectation.* Tahap *Expectation* dilakukan dengan menghitung $\hat{z}_{ig}$ dan $\hat{u}_{ig}$ dengan persamaan :

$$\hat{z}_{ig} = \frac{\hat{\pi}_g f_g(\boldsymbol{x_i}|\boldsymbol{\mu_g}, \Sigma_g, v_g)}{\sum_{g=1}^{G} \hat{\pi}_g f_g(\boldsymbol{x_i}|\boldsymbol{\mu_g}, \Sigma_g, v_g)} \quad (4)$$

$$\hat{u}_{ig} = \frac{v_g + p}{v_g + \delta(\boldsymbol{x_i}, \boldsymbol{\mu_g}|\Sigma_g)} \quad (5)$$

dengan $\hat{z}_{ig}$ adalah probabilitas objek ke-i masuk klaster ke-g dan $\hat{u}_{ig}$ adalah bobot karakteristik (Andrews dan McNicholas, 2012).

Langkah kedua adalah *Conditional Maximization*. Tahap ini menghitung parameter $\hat{\pi}_g, \hat{\boldsymbol{\mu}}_g$ dan $\hat{\Sigma}_g$ dengan rumus sebagai berikut :

$$\hat{\pi}_g = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ig}\hat{u}_{ig} \quad (6)$$

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}\hat{u}_{ig}\, \boldsymbol{x}_i}{\sum_{i=1}^{n} \hat{z}_{ig}\hat{u}_{ig}} \quad (7)$$

$$\hat{\Sigma}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}\, \hat{u}_{ig}(\boldsymbol{x_i} - \boldsymbol{\mu_g})(\boldsymbol{x_i} - \boldsymbol{\mu_g})^T}{\sum_{i=1}^{n} \hat{z}_{ig}} \quad (8)$$

Nilai v_g tidak tersedia dalam bentuk tertutup (*closed form)* dan diperoleh dengan menyelesaikan persamaan (17) :

$$1-\varphi\left(\frac{\hat{v}^{new}}{2}\right)+\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig}\left(\log\hat{u}_{ig}-\hat{u}_{ig}\right)+\log\left(\frac{\hat{v}^{new}}{2}\right)+\varphi\left(\frac{\hat{v}^{old}+p}{2}\right)-\log\left(\frac{\hat{v}^{old}+p}{2}\right)=0 \quad (9)$$

Parameter yang baru akan digunakan kembali pada tahap *Expectation.* Kedua langkah ini akan diulang dan diperbarui secara terus menerus sampai diperoleh nilai yang konvergen. Anggota klaster dikelompokkan menggunakan metode klasifikasi *Maximum a Posteriori* (MAP) sebagai berikut (Agustini, 2017):

$$\text{MAP}\{\hat{z}_{ig}\}=\begin{cases}1, & jika\ \max\{\hat{z}_{ig}\}\ \epsilon\ ke-g\\ 0, & lainnya\end{cases}$$

2.5. *Integrated Completed Likelihood*

Integrated Completed Likelihood dapat digunakan baik untuk pemilihan model maupun jumlah kelompok pada *Model Based Clustering*. ICL dinilai mampu memperkirakan jumlah kelompok dengan stabil dan reliabel. Model terbaik dipilih berdasarkan nilai ICL terbesar (Agustini, 2017).

Rumus ICL didefinisikan sebagaimana persamaan berikut :

$$ICL_g=\ln f\ (y_i)-\frac{p}{2}ln(n) \quad (10)$$

dengan $f(y_i)$ adalah fungsi kepadatan peluang bersama data lengkap, p adalah banyak parameter, dan n adalah banyak observasi.

2.6. Uji Manova

Pengujian perbedaan rata-rata kelompok diperlukan untuk mengetahui adanya perbedaan rata-rata yang berarti di antara klaster yang dihasilkan. Pengujian perbedaan rata-rata pada data multivariat dengan dua atau lebih populasi dapat dilakukan dengan uji Manova *(Multivariate Analysis of Variance)* (Johnson dan Wichern, 2007). Terdapat dua asumsi yang harus dipenuhi pada uji Manova, yaitu normalitas multivariat dan homoskedastisitas. Karakteristik data yang digunakan pada *Model-Based Clustering t-multivariat* adalah terdapat *outlier* pada data *sehingga data tidak memenuhi kedua asumsi uji Manova*.

Ates *et al.* (2019) membandingkan beberapa kriteria pengujian yang digunakan pada uji Manova jika asumsi normalitas dan homoskedastisitas tidak terpenuhi. Untuk distribusi *t* dan asumsi homoskedastisitas tidak terpenuhi, kriteria *Wilk's Lambda* dinilai lebih *robust* dibandingkan kriteria yang lain.

Hipotesis yang digunakan pada uji Manova adalah sebagai berikut.

H_0: $\tau_1=\tau_2=\cdots=\tau_G=0$ (Tidak ada perbedaan antar klaster)

H_1: minimal satu $\tau_i\neq\tau_j$ untuk $i\neq j$ (Ada perbedaan antar klaster)

Kriteria pengujian *Wilk's Lambda* sebagai berikut,

$$\Lambda=\frac{|W|}{|W+B|} \quad (11)$$

dengan $B=\sum_{g=1}^{G}n_g\left(\boldsymbol{x_g}-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x_g}-\overline{\boldsymbol{x}}\right)^T$ dan $W=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{x_{gk}}-\overline{\boldsymbol{x}}_{\boldsymbol{g}}\right)\left(\boldsymbol{x_{gk}}-\overline{\boldsymbol{x}}_{\boldsymbol{g}}\right)^T$.

Statistik Wilks Lambda dapat didekati dengan distribusi F. Keputusan tolak H_0 jika $F_{hitung}>F_{tabel}$ atau $p-value<\alpha(0{,}05)$ yang berarti ada perbedaan antar klaster yang terbentuk (Johnson dan Wichern, 2007).

3. METODE PENELITIAN

Data yang digunakan pada penelitian ini adalah data faktor penyebaran Covid-19 di

Indonesia. Data tersebut meliputi data kepadatan penduduk (X_1) , *testing rate* (X_2), proporsi penduduk usia 50 tahun ke atas (X_3), dan proporsi penduduk yang memiliki kebiasaan cuci tangan (X_4) tiap provinsi. Data diperoleh dari laman resmi Badan Pusat Statistik (BPS) Indonesia dengan tautan https://www.bps.go.id/ tahun 2020 dan laman resmi Data Riset dan Teknologi Covid-19 oleh Kementerian Riset dan Teknologi – Badan Riset dan Inovasi Nasional Republik Indonesia dengan tautan *http://sinta.ristekbrin.go.id/covid/datasets* per tanggal 31 Januari 2021.

Langkah-langkah analisis data yang dilakukan adalah sebagai berikut:

1. Melakukan uji sampel representatif dan uji multikolinearitas
2. Membuat plot dua komponen utama dari data yang digunakan untuk melihat pola dan mengidentifikasi pengelompokkan objek
3. Melakukan uji normal multivariat dan deteksi *outlier* multivariat
4. Melakukan klasterisasi data dengan metode *Model-Based Clustering t-multivariat*
5. Memilih jumlah klaster optimal dengan kriteria *Integrated Completed Likelihood*
6. Melakukan pengujian perbedaan rata-rata menggunakan uji Manova
7. Melakukan interpretasi hasil klasterisasi jika setiap klaster memiliki perbedaan nyata
8. Menarik kesimpulan dan saran

4. HASIL DAN PEMBAHASAN

4.1. Pengujian Asumsi Klaster

Pengujian sampel mewakili populasi (sampel representatif) dilakukan dengan uji *Kaiser-Mayer Olkin* (KMO). Berdasarkan *output* uji KMO dengan *software* R 4.1.0 menggunakan *function "KMOS"* pada *package "REdaS",* diperoleh nilai KMO sebesar 0,5360283. Nilai KMO tersebut lebih besar dari 0,5 sehingga dapat disimpulkan bahwa data sampel yang digunakan merupakan sampel yang cukup representatif.

Pengujian asumsi non-multikolinearitas dilakukan dengan melihat dari nilai VIF masing-masing variabel. Tabel 1. menunjukkan nilai koefisien determinasi beserta nilai VIF dari tiap variabel.

Tabel 1. Nilai VIF Masing-Masing Variabel

Variabel	Koefisien Determinasi	VIF	Keterangan
X_1	0.7641	4,239	VIF ≤ 10
X_2	0.7704	4,355	VIF ≤ 10
X_3	0.1749	1,212	VIF ≤ 10
X_4	0. 2187	1,280	VIF ≤ 10

Berdasarkan Tabel 1, diperoleh nilai VIF ≤ 10 untuk semua variabel. Sehingga dapat disimpulkan tidak terjadi multikolinearitas pada ke-empat variabel atau asumsi non-multikolinearitas terpenuhi.

4.2. Identifikasi Pola Pengelompokkan Data

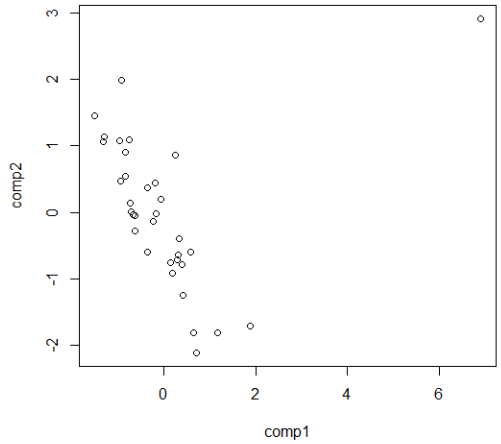
Data faktor penyebaran Covid-19 memiliki skala data yang bervariasi. Oleh karena itu, dilakukan standarisasi data terlebih dahulu sebelum menerapkan analisis komponen utama.

Tabel 2. Tabel Proporsi Varian dari Komponen Utama

Komponen Utama	Proporsi Varian	Proporsi Kumulatif
1	0,5126397	0,5126397
2	0,3140851	0, 8267248
3	0,1421997	0,9689245
4	0,03107548	1,0000000

Berdasarkan Tabel 2, dapat diketahui bahwa komponen utama pertama dan kedua mampu menjelaskan keragaman data cukup tinggi yaitu sebesar 82,67%. Oleh karena itu,

pada penelitian ini akan digunakan skor dua komponen utama untuk mengidentifikasi pola data.

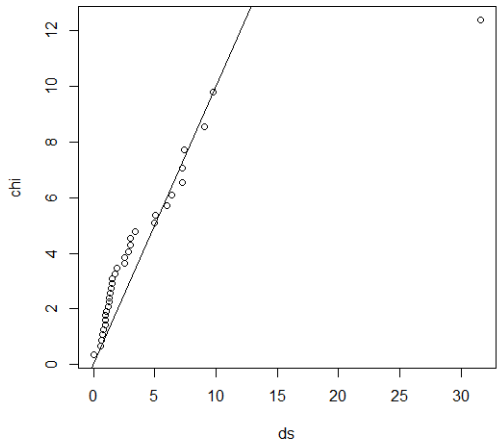


Gambar 1. Plot Dua Komponen Utama

Dari Gambar 1, diketahui bahwa objek pada data cenderung *overlap* (tumpang tindih). Pola penggerombolan data tidak dapat dilihat dengan jelas sehingga jumlah kelompok sulit untuk diidentifikasi.

4.3. Pengujian Asumsi Data

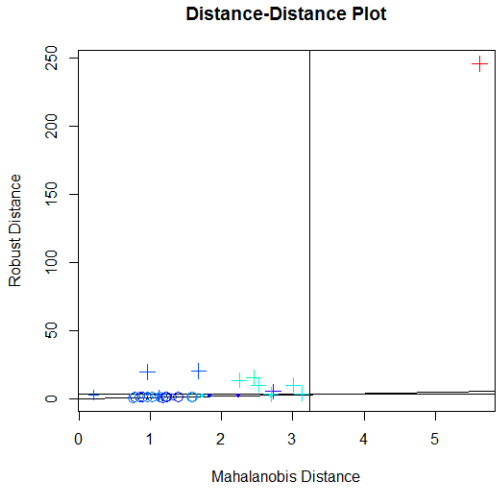
Model-Based Clustering t-Multivariat memiliki asumsi yaitu data tidak berdistribusi normal disebabkan adanya *outlier.*



Gambar 2. Q-Q Plot Data Faktor Penyebaran Covid-19 di Indonesia

Berdasarkan Gambar 2, diketahui bahwa plot tidak membentuk garis lurus sehingga dapat disimpulkan bahwa data tidak mengikuti distribusi normal multivariat.

Pengujian normalitas multivariat secara formal dapat dilakukan dengan menggunakan metode *Kolmogorov-Smirnov*. Perhitungan nilai D pada *software* R 4.1.0, diperoleh nilai $D = 0{,}29176$ dan *p-value* $= 0{,}004656$. Karena nilai $D = 0{,}29176 > W_{(0,95)} = 0{,}227$ dan *p-value* $= 0{,}004656 < \alpha = 0{,}05$ dapat disimpulkan H_0 ditolak yang berarti data faktor penyebaran Covid-19 tidak mengikuti distribusi normal multivariat.



Gambar 3. Diagnostic plot

Gambar 3 menunjukkan *plot* jarak mahalanobis terhadap jarak *robust* yang terbentuk dari data. Berdasarkan gambar 3 dapat diketahui bahwa terdapat pengamatan yang berada pada kuadran I dan II sehingga dapat disimpulkan terdapat *outlier* pada data. Terdapat 9 provinsi yang terdeteksi sebagai *outlier* pada data faktor penyebaran Covid-19, yaitu provinsi Bali, Banten, DI Yogyakarta, DKI Jakarta. Jawa Barat, Jawa Tengah, Jawa Timur, Kep. Riau, dan Sumatera Barat.

4.4. Klasterisasi dengan *Model-Based Clustering t* Multivariat

Klasterisasi faktor penyebaran Covid-19 menggunakan *Model-Based Clustering t*-multivariat dapat dilakukan dengan bantuan *package "teigen"* pada *software* R 4.1.0. Data faktor penyebaran Covid-19 memiliki satuan yang bervariasi. Oleh karena itu, perlu dilakukan standarisasi data. *Package "teigen"* memiliki fungsi *scale* yang secara *default* akan melakukan standarisasi pada data.

Penelitian ini melakukan *running* semua model yang mungkin dari *teigen* dengan banyak kelompok maksimal sebanyak 9 dan menggunakan beberapa inisiasi yaitu "*k-means", "soft",* dan *"hard*". Pemilihan model terbaik dan jumlah kelompok optimal ditentukan melalui nilai ICL terbesar. Berikut adalah tabel yang menunjukkan nilai ICL dari setiap model dan setiap inisiasi untuk jumlah kelompok maksimal adalah 9.

Tabel 3. Jumlah Kelompok Optimal Berdasarkan Inisiasi

Inisiasi	Jumlah kelompok	Nilai ICL
K-means	1	-265.2281
Soft	2	-215.0901
Hard	2	-214.8751

Berdasarkan Tabel 3, inisiasi *hard* menghasilkan nilai ICL terbesar dibandingkan inisiasi *soft dan K-means* yaitu sebesar -214.8751. Oleh karena itu, penelitian ini akan menggunakan inisiasi *hard* dan jumlah klaster optimal sebanyak 2.

Tahap *Expectation* diperoleh vektor nilai $\hat{\boldsymbol{z}}_{\boldsymbol{ig}}$ dan $\hat{\boldsymbol{u}}_{\boldsymbol{ig}}$ sebagai berikut :

$$\hat{\boldsymbol{z}}_{\boldsymbol{ig}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}, \quad \hat{\boldsymbol{u}}_{\boldsymbol{ig}} = \begin{bmatrix} 1.01808 & 0.06261 \\ 0.36350 & 0.11314 \\ 0.16427 & 0.34257 \\ \vdots & \vdots \\ 1.02707 & 0.11097 \end{bmatrix}$$

Tahap *Conditional-Maximization* menghitung nilai parameter $\hat{\pi}_g, \hat{\boldsymbol{\mu}}_g, \hat{\Sigma}_g$ dan v_g berdasarkan nilai $\hat{\boldsymbol{z}}_{\boldsymbol{ig}}$ dan $\hat{\boldsymbol{u}}_{\boldsymbol{ig}}$ yang sudah diestimasi pada tahap sebelumnya. Diperoleh nilai $\hat{\pi}_g$ sebagai berikut :

$$\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{i1}\hat{u}_{i1} = \frac{1}{34}\sum_{i=1}^{n}(1 \times 1.01808) + (0 \times 0.36350) + \cdots + (1 \times 1.02707) = 0.782$$

$$\hat{\pi}_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{i2}\hat{u}_{i2} = \frac{1}{34}\sum_{i=1}^{n}(0 \times 0.06261) + (1 \times 0.11314) + \cdots + (0 \times 0.11097) = 0.218$$

Nilai $\hat{\boldsymbol{\mu}}_g$ dan $\hat{\Sigma}_g$ diperoleh sebagai berikut :

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\hat{u}_{ig}\,\boldsymbol{x}_i}{\sum_{i=1}^{n}\hat{z}_{ig}\hat{u}_{ig}}$$

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} -0.2377764 \\ -0.2094197 \\ 0.2297147 \\ -0.313745 \end{bmatrix}, \qquad \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 0.1127407 \\ -0.1955720 \\ -0.7044711 \\ 1.805061 \end{bmatrix}$$

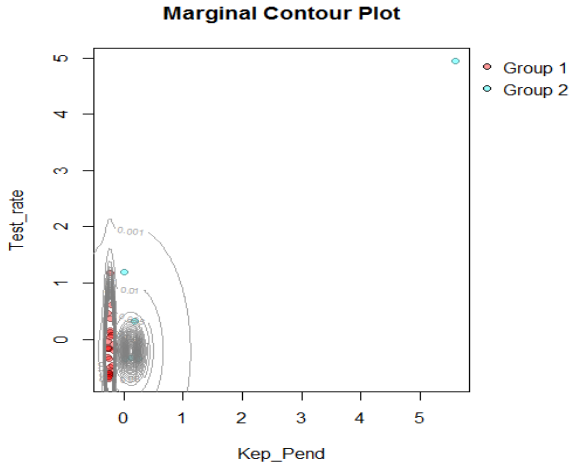
$$\hat{\Sigma}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\,\hat{u}_{ig}(\boldsymbol{x_i} - \boldsymbol{\mu_g})(\boldsymbol{x_i} - \boldsymbol{\mu_g})^T}{\sum_{i=1}^{n}\hat{z}_{ig}}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 0.00077 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.20450 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.90136 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.45827 \end{bmatrix}, \hat{\Sigma}_2 = \begin{bmatrix} 0.02093 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.08842 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.06375 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.54956 \end{bmatrix}$$

Nilai $\boldsymbol{v_g}$ diperoleh dengan menyelesaikan persamaan 18. $\boldsymbol{v_g}$ menyatakan vektor derajat bebas tiap klaster. Berdasarkan Lampiran 13, diperoleh nilai v_g yang berbeda antara klaster 1 dan 2. v_1 dan v_2 berturut-turut bernilai 48.90398 dan 2.00000.

Dengan menggunakan inisiasi *hard* dan jumlah kelompok sebanyak 2, nilai ICL terbesar terdapat pada model CIUU. Model CIUU memiliki bentuk klaster yang merepresentasikan struktur matriks varian kovarian $\Sigma_g = \lambda I A_g {D_g}^T$. Secara keseluruhan, model CIUU memiliki karakteristik yaitu antar klaster memiliki volume yang sama. Akan tetapi kontur fungsi kepadatan dan derajat bebas kedua klaster berbeda. λ_g dengan label C *(Constrained)* yang berarti bahwa kedua kelompok memiliki volume elips yang sama. D_g dengan label I berarti matriks orthogonal *eigenvector* membentuk matriks identitas. A_g dengan label U (*Unconstrained)* berarti kontur yang terbentuk dari kedua kelompok berbeda. ${D_g}^T$ dengan label U (*Unconstrained)* berarti kedua kelompok memiliki derajat bebas berbeda.

Hasil pengelompokkan provinsi di Indonesia berdasarkan faktor penyebaran Covid-19 dapat divisualisasikan dengan melihat *marginal contour plot.*



Gambar 4. *Marginal Contour Plot* Data Penyebaran Covid-19

Berdasarkan gambar 4 dapat dilihat bahwa terdapat irisan antara kedua elips. Hal ini menunjukkan klaster yang terbentuk cenderung *overlap.* Klaster yang *overlap* menunjukkan karakteristik antar kelompok tidak terlalu berbeda.

4.5. Uji Manova

Berdasarkan *output*, diperoleh nilai $F_{hitung} = 9{,}386$ dan *p-value* $= 5.402\text{e} - 05$. H_0 ditolak karena $F_{hitung} = 9{,}386 > F_{(4;29)} = 2{,}701$ dan *p-value* $= 5.402\text{e} - 05 < \alpha = 0{,}05$ sehingga dapat disimpulkan terdapat perbedaan yang berarti antar klaster yang dihasilkan.

4.6. Interpretasi Hasil Pengelompokkan

Hasil pengelompokan provinsi-provinsi di Indonesia berdasarkan data faktor penyebaran Covid-19 menggunakan *Model-Based Clustering t*-Multivariat menghasilkan 2 klaster. Dari 34 provinsi di seluruh Indonesia, terdapat 27 provinsi (79,41%) yang berada pada klaster pertama. Provinsi-provinsi tersebut yaitu provinsi Aceh, Bengkulu, Gorontolo, Jambi, Kalimantan Barat, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Kep. Bangka Belitung, Kep. Riau, Lampung, Maluku, Maluku Utara, Nusa Tenggara Barat, Nusa Tenggara Timur, Papua, Papua Barat, Riau, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Utara, Sumatera Barat, Sumatera Selatan, dan Sumatera Utara. Kelompok kedua beranggotakan 7 provinsi (20,59%) yaitu provinsi Bali, Banten, DI Yogyakarta, DKI Jakarta, Jawa Barat, Jawa Tengah, dan Jawa Timur.

Berdasarkan rata-rata tiap variabelnya, kelompok kedua memiliki rata-rata kepadatan penduduknya, *testing rate*, dan proporsi penduduk berusia lanjut lebih tinggi dibandingkan dengan kelompok pertama. Selain itu, kelompok kedua memiliki proporsi penduduk disiplin mencuci tangan lebih rendah daripada kelompok pertama. Oleh karena itu, secara umum risiko penyebaran Covid-19 pada provinsi di kelompok kedua lebih tinggi dibandingkan provinsi di kelompok pertama.

5. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, diperoleh kesimpulan bahwa klasterisasi provinsi di Indonesia berdasarkan data faktor penyebaran Covid-19 menggunakan *Model-Based Clustering t-Multivariat* menghasilkan 2 klaster. Klaster optimal diperoleh dengan metode *Integrated Completed Likelihood.* Dari 28 model pada *package "teigen",* diperoleh model terbaik yaitu CIUU. Klaster yang terbentuk memiliki volume yang sama, kontur fungsi kepadatan berbeda serta derajat bebas kedua klaster berbeda. Klaster pertama beranggotakan 27 provinsi dan klaster kedua beranggotakan 7 provinsi. Klaster kedua beranggotakan provinsi yang memiliki risiko penyebaran Covid-19 lebih tinggi dibandingkan provinsi pada klaster pertama. Oleh karena itu, pemerintah telah tepat dalam memberlakukan kebijakan khusus pada provinsi di klaster kedua agar tidak terjadi penambahan kasus yang signifikan.

DAFTAR PUSTAKA

Agustini, M. (2017) *Model-Based Clustering dengan Distribusi t Multivariat Menggunakan Kriteria Integrated Completed Likelihood dan Minimum Message Length (Pengelompokan Provinsi di Indonesia Menurut Indikator Pasar Tenaga Kerja Tahun 2012-2015)*. Institut Teknologi Sepuluh Nopember.

Andrews, J. L. and McNicholas, P. D. (2012) 'Model-Based Clustering, Classification, and

Discriminant Analysis via Mixtures of Multivariate t-distributions: The tEIGEN Family', *Statistics and Computing*, 22(5), pp. 1021–1029. doi: 10.1007/s11222-011-9272-x.

Ates, C. *et al.* (2019) 'Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in terms of Type-I Error Rates', *Computational and Mathematical Methods in Medicine*, 2019, p. 8. doi: 10.1155/2019/2173638.

Budiyanti, E. (2020) 'Dampak Virus Corona Terhadap Sektor Perdagangan dan Pariwisata Indonesia', *Info Singkat (Kajian Singkat Terhadap Isu Aktual dan Strategis)*, XII(4), pp. 19–24.

Hair, J. F. *et al.* (2013) *Multivariate Data Analysis*. 7th Editio. Pearson Education Limited.

Johnson, R. A. and Wichern, D. W. and others (2007) *Applied Multivariate Statistics*. 6th Editio. Pearson Education, Inc.

Kaiser, H. F. (1974) 'An index of factorial simplicity', *Psychometrika*, pp. 31–36.

Karo, M. B. (2020) 'Perilaku Hidup Bersih dan Sehat (PHBS) Strategi Pencegahan Penyebaran Virus Covid-19', *Prosiding Seminar Nasional Hardiknas*, pp. 1–4.

Pardede, T. and Prasetyo, B. (2012) *Kajian Metode Berbasis Model Pada Analisis Cluster dengan Perangkat Lunak Mclust*.

Priyadarsini, S. L. and Suresh, M. (2020) 'Factors influencing the epidemiological characteristics of pandemic COVID 19 : A TISM approach Factors in fl uencing the epidemiological characteristics of pandemic COVID 19', *International Journal of Healthcare Management*, 13(2), pp. 89–98. doi: 10.1080/20479700.2020.1755804.

Rosyanti, L. and Hadi, I. (2020) 'Dampak Psikologis dalam Memberikan Perawatan dan Layanan Kesehatan Pasien COVID-19 pada Tenaga Profesional Kesehatan', *Health Information Jurnal Penelitian*, 12(1), pp. 107–130. doi: 10.36990/hijp.vi.191.

Roy, S. and Ghosh, P. (2020) 'Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking', *PLoS ONE*, 15(10), pp. 1–18. doi: 10.1371/journal.pone.0241165.

Velasco, J. M., Tseng, W. and Chang, C. (2021) 'Factors Affecting the Cases and Deaths of COVID-19 Victims', *International Journal of Environmental Research and Public Health*, 18, p. 674. doi: https:// doi.org/10.3390/ijerph18020674.

Yamali, F. R. and Putri, R. N. (2020) 'Dampak Pandemi Covid-19 Terhadap Ekonomi Indonesia', *Ekonomis: Journal of Economics and Business*, 4(2), pp. 384–388. doi: 10.33087/ekonomis.v4i2.179.