Jurnal ICT : Information Communication & Technology
Vol. 21, N0.1, Juli 2022, pp. 1-7
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI:

Interpretasi dan Visualisasi Hasil Clustering Menggunakan K-Medoid untuk Identifikasi Penyebaran Virus Covid-19

Puspita Nurul Sabrina[1*], Ade Kania Ningsih [2], Fatan Kasyidi[3]
[1,2,3]Program Studi Informatika, Universitas Jenderal Achmad Yani, Indonesia
Email: [1]puspita.sabrina@lecture.unjani.ac.id, [2]ade.kanianingsih@lecture.unjani.ac.id,
[3]fatan.kasyidi@lecture.unjani.ac.id

INFORMASI ARTIKEL

Histori artikel:

Naskah masuk, 09 Januari 2022
Direvisi, 20 Januari 2022
Diiterima, 22 Pebruari 2022

Kata Kunci:

Covid19
K-Medoid
Data_Seleksi
Visualisasi_data
Interpretasi

ABSTRAK

Abstract- The spread of the Covid-19 virus in Indonesia continues to occur with an increasing number. How the pattern of the virus spreads needs to be identified to help prevent uncontrolled spread. One way to determine the pattern is to do clustering. The clustering method in this study uses the K-Medoid method, which has been used in several studies on disease analysis. Clustering results need to be processed and analyzed so that knowledge can be captured more easily. This process is called interpretation which can be supported by visualization. The interpretation process is carried out, among others, by visualizing the comparison of 2 attributes that may be related. Another process is processing data that enters clusters by recapitulation/summing by province and other attributes by filtering and selecting records. The interpretation results show that areas with high Population Density, Smaller Areas, High Populations indicate a higher number of cases and new Covid-19 cases. The interpretation for each cluster of virus spread appears that cluster 1 is Big City, Cluster 2 is City of Tourism and Neighbors with other countries, Cluster 3 is Java Island and Cluster 4 is other cities outside the other three clusters.

Abstrak- Penyebaran virus Covid-19 di Indonesia terus terjadi dengan jumlah yang terus meningkat. Bagaimana pola penyebaran virus perlu diidentifikasi untuk membantu pencegahan penyebaran yang tidak terkendali. Salah satu cara menentukan pola adalah dengan melakukan *clustering*. Metode *clustering* dalam penelitian ini menggunakan metode K-Medoid, yang mana metode ini telah digunakan dalam beberapa penelitian tentang analisa penyakit. Hasil *Clustering* perlu diolah dan dianalisis agar pengetahuan dapat ditangkap lebih mudah. Proses ini disebut interpretasi yang dapat didukung dengan visualisasinya. Proses interpretasi dilakukan di antaranya dengan melakukan visualisasi perbandingan 2 atribut yang mungkin terkait. Proses lainnya dengan pengolahan data yang masuk klaster-klaster dengan rekapitulasi/penjumlahan menurut provinsi dan atribut lain dengan cara *filtering* dan seleksi *record*. Hasil interpretasi diperoleh kesimpulan daerah dengan Kepadatan Penduduk tinggi, Area Wilayah yang lebih kecil, Populasi yang tinggi menunjukkan jumlah kasus dan kasus baru Covid-19 lebih tinggi. Interpretasi untuk setiap klaster penyebaran virus tampak klaster 1 adalah Kota Besar, Klaster 2 adalah Kota Wisata dan Bertetangga dengan negara lain, Klaster 3 adalah Pulau Jawa dan Klaster ke 4 adalah kota-kota lain diluar ketiga klaster lainnya.

Copyright © 2021 LPPM - STMIK IKMI Cirebon
This is an open access article under the CC-BY license

***Penulis Korespondensi**:*

Puspita Nurul Sabrina
Program Studi Informatika
Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman, Kota Cimahi, Indonesia
Email: puspita.sabrina@lecture.unjani.ac.id

1. Pendahuluan

Penyebaran penyakit akibat virus yang semakin meluas menimbulkan keresahan. Pola penyebaran virus memungkinkan untuk diidentifikasi berdasarkan data penderita penyakit. Salah satu cara menentukan pola adalah dengan melakukan *clustering* berdasarkan data penderita dari berbagai wilayah. Salah satu virus yang saat ini berkembang penyebarannya adalah virus corona. Penyebaran virus ini yang sangat cepat perlu segera ditangani agar jumlah penderita tidak melonjak tinggi. Untuk mengantisipasi lonjakan dapat dengan melakukan Analisa penyebaran yang sudah terjadi. Salah satu proses Analisa dapat berdasarkan pola penyebaran dalam klaster-klaster tertentu. Salah sat teknik dalam membentuk kelompok klaster ini adalah metoda data mining *clustering*. Berbagai metode *clustering* sudah berkembang, salah satunya metode K-Medoid yang mana metode ini telah digunakan dalam beberapa penelitian tentang analisa penyakit.

Proses *clustering* adalah dengan melakukan mining pada sejumlah dataset yang berkaitan dengan Covid-19. Pada penelitian ini dataset diperoleh dari dari open data penderita virus corona. Data ini dengan metode data mining *clustering* K-Medoid akan diproses untuk membantu mengidentifikasi berbagai faktor terkait penyebaran virus.

Jumlah penderita yang terus bertambah di Indonesia mengerucutkan penelitian ini dengan dataset penderita virus yang berada di Indonesi yang tersebar dalam 34 provinsi di Indonesia. Identifikasi pola penyebaran menjadi hal yang dibutuhkan untuk mengantisipasi penyebaran yang lebih luas dan lebih banyak.

Pada penelitian Segmentasi Leaf menggunakan Algoritma K-Means dan K-Medoid untuk Identifikasi Penyakit yang merupakan algoritma pengelompokan yang paling umum digunakan sebagai pendekatan dasar. Algoritma pengelompokan K-means dan K-Medoids diperiksa dan dianalisis, dari hasil, ditemukan bahwa K-berarti berkinerja baik dalam situasi ketika penyakit ini mudah dibagi dari *leaf* karena menghitung jarak berdasarkan nilai rata-rata[1].

Myocardial Infarction (MI) adalah penyakit yang sangat umum di dunia. Tujuan utama dari penelitian Myocardial Infarction (MI) adalah untuk memprediksi infark miokard menggunakan teknik *clustering* data mining. Sistem ini dapat mengenali dan memilih intelijen tersembunyi dari set data historis infark miokard. Makalah ini telah memeriksa prediksi serangan jantung dengan jumlah yang lebih banyak atribut input. Rumah sakit mengeksploitasi istilah medis seperti usia, jenis kelamin, tekanan darah, kolesterol, dan sebagainya untuk memprediksi pasien yang mendapatkan penyakit jantung. Seleksi Fitur dalam penambangan data adalah pendekatan untuk meminimalkan jumlah atribut input yang diabaikan. Awalnya jumlah atribut adalah 14 dan setelah pemilihan fitur, jumlah atribut dikurangi menjadi 8 atribut [2].

Metode baru untuk mengidentifikasi node yang berpengaruh dalam jaringan yang kompleks dengan struktur komunitas diusulkan. Metode ini menggunakan probabilitas transfer informasi antara setiap pasangan node dan algoritma *clustering* k-medoid. Hasil percobaan menunjukkan bahwa node berpengaruh yang diidentifikasi oleh metode k-medoid dapat mempengaruhi ruang lingkup yang lebih besar dalam jaringan dengan struktur komunitas yang jelas daripada algoritma greedy tanpa mengurangi jumlah yang diharapkan dari node yang dipengaruhi [3].

Penyakit Ginjal Kronis ("CKD") dan penyakit penyerta, diabetes, hipertensi dan penyakit kardiovaskular ("CVD"), sering diukur dengan prosedur rutin dan tes laboratorium, yang membuat sejumlah besar data historis tentang populasi pasien ini. Penelitian dilakukan untuk studi retrospektif berdasarkan data Electronic Health Records ("EHR"), untuk mengidentifikasi pola dalam pengembangan CKD [4].

Penelitian sebelumnya menyarankan pembentukan klaster gejala yang diimplementasikan melalui aplikasi praktis, seperti perumusan intervensi terapeutik yang lebih efektif yang membahas efek gabungan gejala daripada mengobati masing-masing gejala secara terpisah. Sebagian besar penelitian yang telah berusaha mengidentifikasi kelompok yang selamat dari kanker payudara mengandalkan studi penelitian tradisional. Penelitian ini berupaya untuk menentukan pola cluster gejala pada penderita kanker payudara yang berasal dari media sosial dan data penelitian dengan menggunakan pengelompokan K-Medoid yang ditingkatkan [5].

Pada awal maret Indonesia sedang dilanda masuknya wabah virus Covid-19. Setiap hari kasus penyebaran covid-19 di Indonesia terus meningkat. Masyarakat diminta untuk melakukan *social distancing* guna mamutus rantai penyebaran Covid-19 yang tersebar diberbagai wilayah.di Indonesia. Data yang telah ditampung banyak sekali, dari data tersebut dapat dilihat pola–pola penentuan pengelompokan penyebaran covid-19 dilakukan berdasarkan nilai tes. Penelitian ini menggunakan metode K-Medoids agar dapat diketahui pola pemilihan penentuan pengelompokan penyebaran Covid-19 bagi masyarakat [6].

Hasil *clustering* yang berkualitas baik harus didukung dataset yang disiapkan juga harus berkualitas baik. Tahap *preprosesing* dan *data selection* diperlukan untuk itu. Pembersihan data (untuk membuang data yang tidak konsisten dan *noise*) dan menjadi dasar persiapan dan pengelolaan dataset [7].

Pengelompokan pasien Covid-19 menggunakan metode K-Means dan K-Medoid berdasarkan atribut usia, jenis penularan, jenis kelamin, faskes dan kecamatan. Hasil penelitian ini menunjukkan algoritma K-Means lebih optimal dibanding K-Medoid pada pengelompokkan pasien Covid-19 khususnya di Kota Dumai. Hal ini dibuktikan dengan nilai terbaik DBI K-Means sebesar 0,139 dengan k = 4 [8].

K-Medoids yang merupakan metode analisis dapat secara partisional *clustering* yang bertujuan untuk mendapatkan suatu set k-cluster di antara data yang paling mendekati suatu objek dalam pengelompokan kumpulan data. Hasil penelitian berdasarkan pengelompokan penyebaran covid-19 menunjukkan bahwa masyarakat yang berasal dari berbagai wilayah di Indonesia terdampak covid. Ciri-ciri dengan suhu badan di atas 36,9◦ c dan dengan disertai demam dan batuk berkelanjutan menunjukkan salah satu ciri-ciri gejalah covid-19 [6].

Penelitian lain mempromosikan pemahaman tentang demografi dan variabel sosial ekonomi (*spasial non-stasioner* yang tidak teramati) yang mempengaruhi pola spasial COVID-19 di Oman. Pemahaman ini untuk menjadi dasar memilih strategi yang lebih tepat dalam mengatasi pandemi di masa depan dan juga untuk mengalokasikan pencegahan yang lebih efektif [9].

Berdasarkan data peta sebaran Covid-19 dan regulasi tentang aturan protokol kesehatan dari berbagai sumber data online baik secara terstruktur ataupun tidak terstruktur menggunakan *cleansing data Business Intelligence* (BI) ditemukan tiga pola penyebaran virus Corona yaitu Kasus Tinggi, Kasus Sedang dan Kasus Rendah dari 34 Provinsi se Indoensia. Klaster menunjukan Provinsi Jawa Barat menjadi urutan pertama kasus tertinggi selama Mei 2021 dengan rata-rata kasus 21% kemudian dilanjut DKI Jakarta 14% [10].

Analisis klaster multivariat dari indeks sosial ekonomi dilakukan untuk mengidentifikasi daerah-daerah yang memiliki kerawanan sosial serupa. Hasilnya berupa rangkaian peta jarak efektif, kemungkinan wabah, kapasitas rumah sakit dan kerentanan social Utara dan Timur Laut dengan risiko tinggi wabah COVID-19 yang juga sangat rentan secara social [11].

Generalized Additive Models (GAM) digunakan untuk memodelkan kurva kumulatif dan harian untuk kasus dan kematian yang dikonfirmasi Penggunaan fungsi GAM untuk memprediksi kasus dan kematian yang dikonfirmasi terbukti memadai, bahkan dengan terjadinya lonjakan atau "gelombang kedua" dalam rangkaian ini. pendekatan baru bahkan memungkinkan identifikasi beberapa titik belok di seluruh rangkaian kasus dan kematian yang dikonfirmasi setiap hari: suatu kemajuan jika dibandingkan dengan fungsi pertumbuhan tradisional [12].

Identifikasi penyebaran Covid-19 berdasarkan dataset dengan metode *clustering* K-Medoid dilakukan sehingga dapat diperoleh pengetahuan tentang tentang faktor apa saja yang mempengaruhi penyebaran virus, khususnya virus Covid-19. Permasalahan yang dihadapi adalah dalam memahami hasil klaster. Perlu Analisa dalam memahami dan menangkap pengetahuan dalam hasil klaster yaitu melalui proses interpretasi. Selain itu karena hasil klaster bisa menyangkut atribut data yang beragam dengan jumlah record yang besar, perlu visualisasi yang representatif dalam menginterpretasikan hasil *clustering* dan pengetahuan yang diperoleh.

2. Metode

Sumber dataset yang digunakan adalah Dataset pandemi Covid-19 di Indonesia dengan timeseries. Dataset ini merupakan kompilasi dari berbagai sumber data terbuka, antara lain: covid19.go.id (data pandemi), kemendagri.go.id (data demografi), bps.go.id (data demografis), serta beberapa perhitungan hubungan antar data. Dataset ini berisi rangkaian waktu kejadian pandemi Covid-19 di Indonesia, di tingkat negara hingga provinsi. Data set sebanyak 8490 record. Data dimulai dari tanggal 1 Maret 2020 hingga 19 November 2020.

Adapun atrribut dataset terdiri dari 36 atribut yaitu Date, Location ISO Code,Location, New Cases, New Deaths, New Recovered,New Active Cases, Total Cases, Total Deaths,Total Recovered, Total Active Cases, Location Level,City or Regency, Province, Country, Continent, Island, Time Zone, Special Status, Total Regencies, Total Cities, Total Districts, Total Urban Villages, Total Rural Villages, Area (km^2), Population, Population Density, Longitude, Latitude, New Cases per Million, Total Cases per Million, New Deaths per Million, Total Deaths per Million, Case Fatality Rate, Case Recovered Rate, Growth Factor of New Cases, Growth Factor of New Deaths.

Pada penelitian ini, berdasarkan kondisi dataset dan tujuan *clustering* untuk identifikasi penyebaran virus, maka tahapan penelitian adalah sebagai berikut :

1. Persiapan Data
   Pembersihan dan Integrasi data yaitu untuk membuang data yang tidak konsisten dan noise. Selanjutnya Seleksi dan Transformasi data (data diubah menjadi bentuk yang sesuai untuk proses mining).
2. Proses Mining, proses ekstraksi pola dari data yang ada.
3. Eksperimen beberapa kombinasi atribut data yang digunakan
4. Presentasi pengetahuan dengan teknik visualisasi.
5. Proses interprestasi pola yang dihasilkan dan visualisasnyai dengan Teknik filtering dan

seleksi untuk identifikasi pola penyebaran virus sehingga diperoleh pengetahuan.

Pada bagian ini membahas metode yang dilaksanakan dalam penelitian, dimulai dari Persiapan Data hingga Interpretasi dan Visualisasi.

1. Persiapan Data

Persiapan dan pengelolaan Dataset dengan Data Selection, Preprocessing, Transformation. Data Selection merupakan proses mengoptimalkan jumlah data yang digunakan untuk proses mining dengan tetap merepresentasikan data aslinya.

a. Pembersihan data

Kondisi awal data set yang beberapa value terdiri dari nilai NaN atau Null, outliers, isu encoding data. Isu encoding diselesaikan dengan penyesuaian dengan format yang relevan sebagai input untuk masuk ke mesin data mining. Data dibersihkan karena kondisi di atas sehingga sekitar 3% data dibuang. Pembersihan dari nilai negatif yang tidak relevan dengan nilai minimum dari atribut New_Active_Cases> dan Total_Active_Cases.

b. Seleksi dan Transformasi data

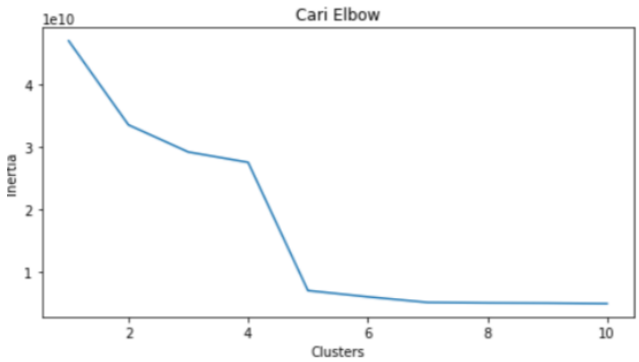
Beberapa field yang non-numerical tidak dimasukkan dalam perhitungan yaitu Date, Location_ISO_Code,Location_Level City_or_Regency, Province, Country, Continent, Island, Time_Zone, Special_Status,Case_Fatality_Rate,Case_Recovered_Rate.

Beberapa atribut diseleksi untuk mendeskripsikan data untuk kebutuhan yang berkaitan distribusi dataset. Distribusi data set dilihat berdasarkan atribut Date, Location_ISO_Code,Location_Level City_or_Regency, Province, Country, Continent, Island.

2. Proses Mining dengan K-Medoid

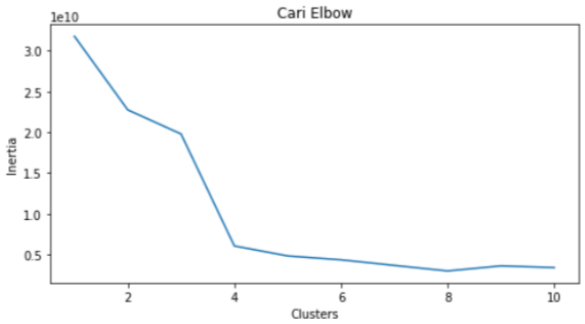
Metode Klaster dengan metode K-Medoid, dengan melalui perhitungan jumlah klaster dengan metode elbow. Hasil metode elbow untuk K-Medoid diperikasa dengan dua kondisi, sebelum preprosessing dan setelah preprosesing.

Hasil elbow sebelum tahap preposessing seperti terlihat pada gambar 1, menghasilkan nilai klaster yaitu 5 klaster.



Gambar 1. Elbow sebelum preprocessing

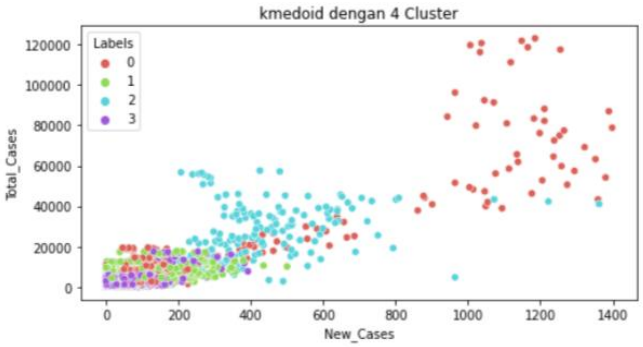
Hasil elbow setelah tahap preposessing seperti terlihat pada gambar 2, menghasilkan nilai klaster 4 klaster.



Gambar 2. Elbow setelah preprocessing

Metode K-Medoid dijalankan dan membentuk 4 klaster, dengan atribut New_Cases,New_Deaths, New_Recovered, New_Active_Cases, Total_Cases, Total_Deaths, Total_Recovered, Total_Active_Cases, Area_(km2), Population, Population_Density, Longitude, Latitude, Growth_Factor_of_New_Cases, Growth_Factor_of_New_Deaths.

Hasil *clustering* dengan visualisasi terhadap kasus baru dan total kasus dapat dilihat pada gambar 3 yang mana tersebar dalam empat klaster.



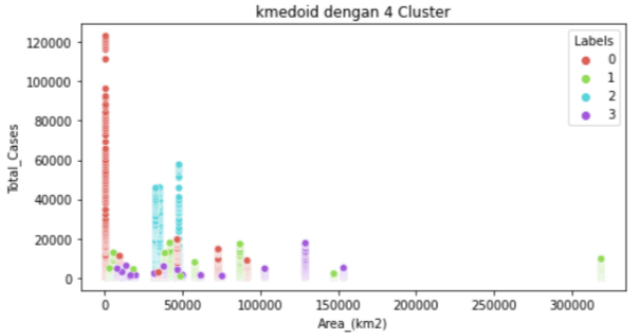
Gambar 3. Visualisasi Hasil Klaster

3. Eksperimen beberapa kombinasi atribut data yang untuk mendapat pengetahuan adalah sebgai berikut:
   a) Eksperimen dengan melihat penyebaran total kasus Covid terhadap Area, dengan tujuan melihat keterkaitan antara jumlah total kasus baru yang tersebar dalam area tertentu.
   b) Eksperimen dengan melihat penyebaran penambahan kasus baru Covid terhadap Area dengan tujuan melihat keterkaitan antara penambahan total kasus baru yang tersebar dalam area tertentu.
   c) Eksperimen dengan melihat penyebaran kasus baru dan Kepadatan Penduduk dengan tujuan melihat keterkaitan antara penambahan total kasus baru dengan tingkat kepadatan penduduk.
   d) Eksperimen dengan melihat penyebaran kasus baru dan Populasi dengan tujuan

melihat keterkaitan antara penambahan total kasus baru dengan jumlah Populasi Penduduk.

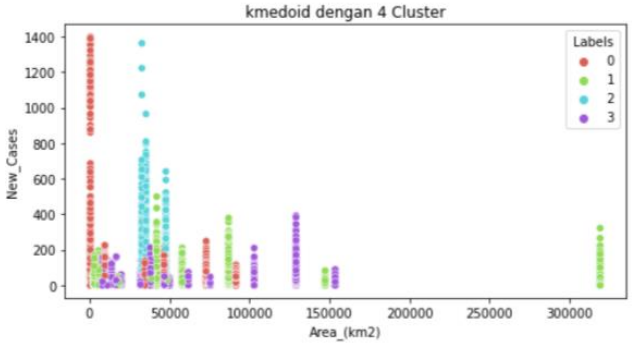
e) Evaluasi pola yang ditemukan dari hasil eksperimen.

4. Presentasi pengetahuan dengan teknik visualisasi.
   Hasil *clustering* dengan visualisasi terhadap total kasus dan Area dapat dilihat pada gambar 4.



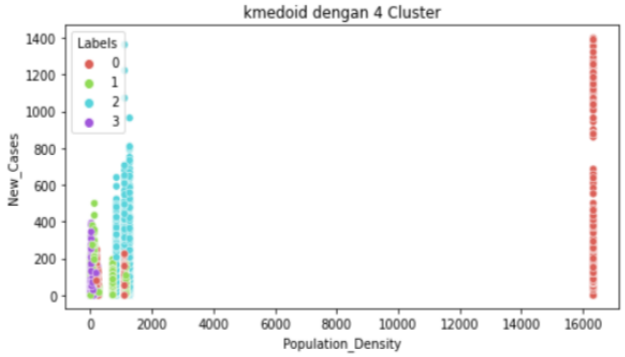
Gambar 4. Visualisasi terhadap total kasus dan Area

Hasil *clustering* dengan visualisasi terhadap kasus baru dan Area dapat dilihat pada gambar 5.



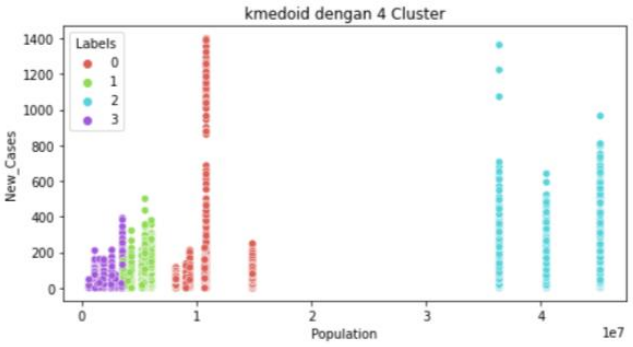
Gambar 5. Visualisasi terhadap kasus baru dan Area

Hasil *clustering* dengan visualisasi terhadap kasus baru dan Kepadatan Penduduk dapat dilihat pada gambar 6.



Gambar 6. Visualisasi terhadap kasus baru dan Kepadatan Penduduk

Hasil *clustering* dengan visualisasi terhadap kasus baru dan Populasi dapat dilihat pada gambar 7.



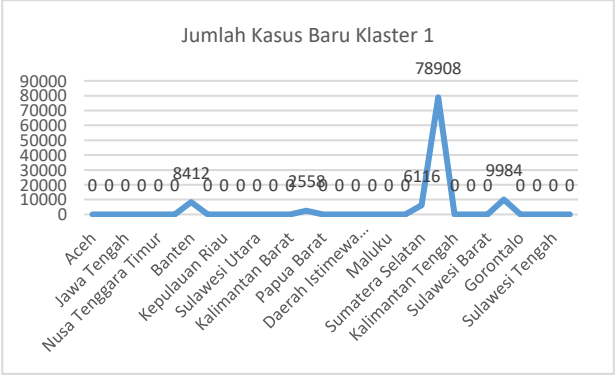
Gambar 7. Visualisasi terhadap kasus baru dan Populasi

5. Proses interpretasi pola yang dihasilkan dengan dukungan visualisasnyai.

Interpretasi pola yang ditemukan dari hasil eksperimen dilakukan dengan Teknik filtering dan seleksi untuk identifikas pola penyebaran virus sehingga diperoleh pengetahuan. Pertama, analisis dilakukan pada hasil eksperimen dengan membandingkan atribut-atribut yang terdapat dalam dataset. Menampilkan hasil analisis dan pengolahan data hasil klaster dalam bentuk grafik. Interpretasi dilakukan dengan melihat sejumlah atribute penting seperti jumlah kasus, jumlah kasus baru, area/wilayah, kepadatan penduduk dll pada setiap klaster untuk dibandingkan penyebarannya pada 33 provinsi di Indonesia.

Interpretasi dengan melakukan pengolahan data hasil *clustering*. Setiap klaster dihitung jumlah kasus baru covid-19 untuk setiap propinsi yang masuk dalam klaster tersebut. Jumlah penderita selama jangka waktu tersebut dijumlahkan berdasarkan propinsi. Total yang muncul dianalisis berdasarkan karakter wilayahnya.

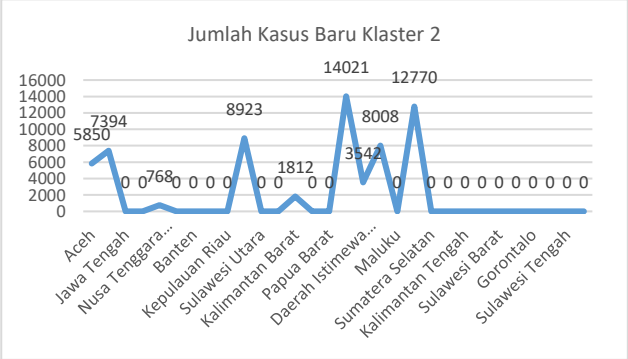
Berikut ini hasil perhitungan dan visualisasi penambahan Jumlah Kasus Baru pada Klaster 1 (Gambar 8).



Gambar 8. Visualisasi Interpretasi Klaster 1

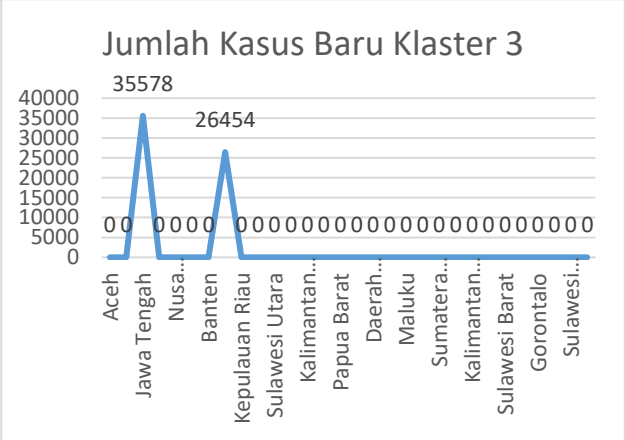
Tampak klaster 1 didominasi oleh penambahan kasus pada provinsi Banten, Lampung, Sumatera Selatan, DKI Jakarta dan Sumatera Utara.

Berikut ini hasil interpretasi dan visualisasi penambahan Jumlah Kasus Baru pada Klaster 2 (Gambar 9).



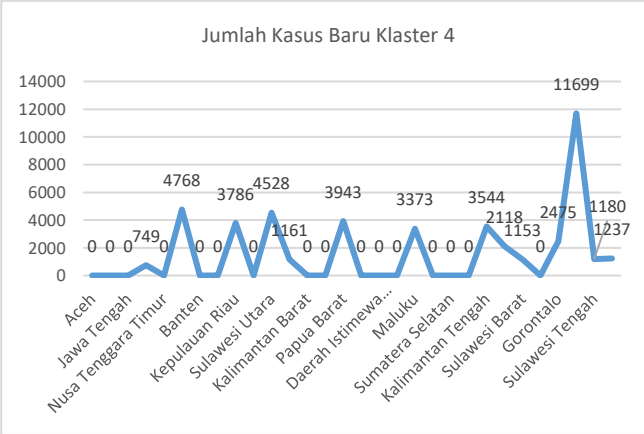
Gambar 9. Visualisasi Interpretasi Klaster 2

Tampak klaster 2 penambahan kasus baru didominasi pada provinsi Aceh, Bali, Nusa Tenggara Timur, Papua, Kalimantan Barat, Sumatera Barat, DI Yogyakarta, Kalimanta Selatan, Riau.
Berikut ini hasil interpretasi dan visualisasi penambahan Jumlah Kasus Baru pada Klaster 3 (Gambar 10).



Gambar 10. Visualisasi Interpretasi Klaster 3

Tampak klaster 3 penambahan kasus baru didominasi pada daerah Jawa Tengah dan Jawa Timur.
Berikut ini hasil interpretasi dan visualisasi penambahan Jumlah Kasus Baru pada Klaster 4 (Gambar 11).



Gambar 11. Visualisasi Interpretasi Klaster 4

Tampak klaster 4 penambahan kasus baru didominasi pada daerah Bangka Belitung, Sulawesi Tenggara, Kepulauan Riau, Sulawesi Utara, Bengkulu , Papua Barat, Maluku, Kalimantan Tengah, Maluku Utara, Kalimantan Timur, Sulawesi Tengah.

3. Hasil dan Diskusi

Eksperimen dengan melihat penyebaran total kasus Covid terhadap Area, dengan tujuan melihat keterkaitan antara jumlah total kasus Covid-19 yang tersebar dalam area tertentu menunjukkan bahwa area yang lebih kecil antara 0 – 50.000 km^2, jumlah penderita Covid paling banyak dibanding Area yang lebih luas di atas 50.000 km^2.

Percobaan dengan melihat penyebaran penambahan kasus baru Covid terhadap Area dengan tujuan melihat keterkaitan antara penambahan total kasus baru yang tersebar dalam area tertentu menunjukkan hal yang mirip, bahwa area yang lebih kecil antara 0 – 50.000 km^2, jumlah penambahan penderita Covid bertambah lebih besar dibanding Area yang lebih luas di atas 50.000 km^2.

Eksperimen dengan membandingkan atribute terkait penyebaran kasus baru dan Kepadatan Penduduk dengan tujuan melihat keterkaitan antara penambahan total kasus baru dengan tingkat kepadatan penduduk menunjukkan kepadatan yang paling tinggi 160.000 adalah yang kasus baru tinggi, diikuti dengan kepadatan 0 – 20.000.

Eksperimen dengan melihat penyebaran kasus baru dan Populasi dengan tujuan melihat keterkaitan antara penambahan total kasus baru dengan jumlah Populasi Penduduk 1 juta dan 4 juta keatas menunjukkan jumlah kasus baru yang tinggi.

Selanjutnya dari sisi klaster yang terbentuk diperoleh hasil interpretasi tampak klaster 1 didominasi oleh penambahan kasus pada provinsi Banten, Lampung, Sumatera Selatan, DKI Jakarta dan Sumatera Utara. Jika diperhatikan ini merupakan provinsi yang terdapat kota besar. Sementara jika bukan kota besar seperti lampung namun berada bersebelahan dan dekat dengan kota besar.

Hasil interpretasi tampak klaster 2 penambahan kasus baru didominasi pada daerah Jawa Tengah dan Jawa Timur. Merupakan provinsi di pulau jawa, yang kemungkinan terpapar karena bertetangga dengan kota besar dan kota wisata.

Hasil interpretasi tampak klaster 3 penambahan kasus baru didominasi pada daerah Jawa Tengah dan Jawa Timur. Merupakan provinsi di pulau jawa, yang kemungkinan terpapar karena bertetanggan dengan kota besar dan kota wisata.

Sementara tampak klaster 4 penambahan kasus baru didominasi pada daerah Bangka Belitung, Sulawesi Tenggara, Kepulauan Riau, Sulawesi Utara, Bengkulu, Papua Barat, Maluku, Kalimantan Tengah, Maluku Utara, Kalimantan Timur, Sulawesi Tengah. Klaster ini kemungkinan menunjukkan kota-kota lain di luar kota besan dan kota wisata dimana terimbas covid dari klaster lainnya.

4. Kesimpulan

Dataset Covid-19 yang diujikan dalam penelitian ini terdiri dari 36 atribut, karena itu perlu untuk dilakukan seleksi yang optimal agar hasil *clustering* lebih baik. Dengan menggunakan ke-36 atribut dan mengurangi delapan atribut string hasilnya tidak bagus dalam *clustering*. Begitu pula dengan hasil dalam metode Elbow. Selanjutnya akan dilakukan seleksi atribut kembali untuk menghasilkan klaster yg lebih optimal. Klaster yang dihasilkan yaitu 4 cluster. Interpretasi dilakukan dengan melakukan analisis data dari penyebaran yang dihasilkan dari setiap klaster. Dengan membandingkan atribut saat proses *mining* dapat diperoleh bahwa daerah dengan kepadatan tenduduk tinggi, wilayah yang lebih kecil, Populasi yang tinggi menunjukan jumlah kasus dan kasus baru yang lebih tinggi.

Proses rekapitulasi per provinsi untuk setiap data yang tersebar pada klaster dapat mendukung interpretasi untuk melihat pola penyebaran covid yang dilihat dari total penambahan kasus untuk setiap provinsi di Indonesia. Tampak klaster 1 adalah Kota Besar, Klaster 2 adalah Kota Wisata dan Bertetangga dengan Negara lain, Klaster 3 adalah Pulau Jawa dan Klaster ke 4 adalah kota-kota lain diluar ketiga klaster.

Daftar Pustaka

[1] S. K. Muruganandham, D. Sobya, S. Nallusamy, D. K. Mandal, and P. S. Chakraborty, "Study on Leaf Segmentation Using K-Means and K-Medoid Clustering Algorithm for Identification of Disease," *Indian Journal of Public Health Research & Development*, vol. 9, no. 5, p. 289, 2018, doi: 10.5958/0976-5506.2018.00456.4.

[2] M. Umamaheswari and P. I. Devi, "Prediction of myocardial infarction using K-medoid clustering algorithm," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Mar. 2017, pp. 1–6. doi: 10.1109/ITCOSP.2017.8303128.

[3] X. Zhang, J. Zhu, Q. Wang, and H. Zhao, "Identifying influential nodes in complex networks with community structure," *Knowledge-Based Systems*, vol. 42, pp. 74–84, Apr. 2013, doi: 10.1016/j.knosys.2013.01.017.

[4] M. Lenart, N. Mascarenhas, R. Xiong, and A. Flower, "Identifying risk of progression for patients with Chronic Kidney Disease using clustering models," in *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)*, Apr. 2016, pp. 221–226. doi: 10.1109/SIEDS.2016.7489303.

[5] Q. Ping, C. C. Yang, S. A. Marshall, N. E. Avis, and E. H. Ip, "Breast Cancer Symptom Clusters Derived From Social Media and Research Study Data Using Improved K -Medoid Clustering," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 63–74, Jun. 2016, doi: 10.1109/TCSS.2016.2615850.

[6] S. Sindi *et al.*, "ANALISIS ALGORITMA K-MEDOIDS CLUSTERING DALAM PENGELOMPOKAN PENYEBARAN COVID-19 DI INDONESIA," *Jurnal Teknologi Informasi*, vol. 4, no. 1, 2020.

[7] J. E. C. Saire, "Data Mining Approach to Analyze Covid19 Dataset of Brazilian Patients," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.11344

[8] U. Rizkya Gurning, "Penerapan Algoritma K-Means dan K-Medoid untuk Pengelompokkan Data Pasien Covid-19," *Technology and Science (BITS)*, vol. 3, no. 1, 2021, doi: 10.47065/bits.v3i1.1003.

[9] K. M. al Kindi *et al.*, "Demographic and socioeconomic determinants of COVID-19 across oman-a geospatial modelling approach," *Geospatial Health*, vol. 16, no. 1, pp. 145–160, 2021, doi: 10.4081/gh.2021.985.

[10] A. S. Sinaga, A. S. Sitio, R. Ramadhani, and A. M. Karimah, "Analisa Big Data Penyebaran Covid-19 Berdasarkan Peta Sebaran dan Peraturan Protokol Dengan Business Intelligence (BI)," *Jurnal Ilmiah Komputasi*, vol. 20, no. 3, Sep. 2021, doi: 10.32409/jikstik.20.3.2775.

[11] J. E. C. Saire, "Data Mining Approach to Analyze Covid19 Dataset of Brazilian Patients," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.11344

[12] A. M. B. de Oliveira, J. M. Binner, A. Mandal, L. Kelly, and G. J. Power, "Using GAM functions and Markov-Switching models in an evaluation framework to assess countries' performance in controlling the COVID-19 pandemic," *BMC Public Health*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12889-021-11891-6.