MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

# Analisis Algoritma K-Means dalam Pengelompokkan Persebaran Covid-19 di Indonesia

**[1)] Nurul Khasanah Fitriyani**
Universitas AMIKOM Yogyakarta, Jl. RIngroad Utara, Yogyakarta, Indonesia
E-Mail: nurul.fitriyani@students.amikom.ac.id

**[2)] Ferian Fauzi Abdulloh**
Universitas AMIKOM Yogyakarta, Jl. RIngroad Utara, Yogyakarta, Indonesia
E-Mail: ferian.fauzi@amikom.ac.id

**ABSTRACT**
Covid-19 or Coronavirus is a virus that is found in humans and animals. This virus can infect humans to cause various diseases such as flu, to serious diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). In Indonesia, the spread of Covid-19 cases continues to increase and is evenly distributed in all provinces in Indonesia because of the fairly rapid spread due to the vast area in Indonesia, making it possible for grouping based on regions in Indonesia to be needed which will result in the center points of the spread of this Covid-19 case. This study aims to group Covid-19 data into a cluster using the K-Means Clustering Data Mining Algorithm. The Covid-19 data used in this study is Covid-19 data on July 6, 2021 which was taken from the official website of Kawal Covid-19 (KawalCovid-19.id). The attributes used are positive cases, recovered, and died. The clusters formed from the results of research using K-Means Clustering are 3 clusters with the first cluster consisting of 2 provinces, the second cluster 3 provinces, and for the third cluster 29 provinces. The cluster with the largest Covid-19 spread rate is cluster one. From this study, the accuracy was 91.176% and evaluated using the Davies-Bouldin Index yielded a fairly good cluster result with a value of 0.493371469.

**Keyword : Covid-19, Data Mining, K-Means, Indonesia**

## PENDAHULUAN

Covid-19 adalah virus penyakit yang menyebabkan gangguan pada infeksi saluran pernapasan manusia, dari flu biasa hingga penyakit yang serius seperti Middle East Respiratory Syndrome (MERS) dan Severe Acute Respiratory Syndrome (SARS) [1]. Gejala dari penderita Covid-19 ini mirip dengan SARS tetapi angka kematian SARS (9,6%) lebih tinggi dibandingkan dengan Covid-19 (kurang dari 5%). Namun, untuk jumlah kasus Covid-19 jauh lebih besar dan memiliki penyebaran yang lebih luas dan cepat apabila dibandingkan dengan SARS [2].

Luasnya wilayah di Indonesia saat ini memungkinkan diperlukannya pengelompokan berdasarkan wilayah di Indonesia yang akan menghasilkan titik-titik pusat penyebaran dari kasus Covid-19. Clustering atau pengelompokan tidak memiliki variable target dalam melakukan suatu pengelompokan pada proses clustering. Clustering adalah suatu proses mengelompokkan kelas yang mempunyai kesamaan objek [3].

Penelitian ini menggunakan algoritma K-Means Clustering agar dapat diketahui pola penentuan pengelompokan persebaran covid-19 di Indonesia. K-Means merupakan salah satu algoritma clustering yang masuk dalam unsupervised learning untuk mengelompokkan data menjadi beberapa kelompok dengan beberapa cluster [4]. Dapat diketahui beberapa kebijakan pemerintah berbagai strategi seperti promotif, preventif dan kuratif untuk menekan penyebaran kasus Covid-19 telah dilakukan sehingga direalisasikan adanya PSBB dan PSBL serta new normal life agar penyebaran kasus Covid-19 ini tidak melonjak tinggi [5]. Dengan menggunakan algoritma K-Means Clustering ini sehingga dapat diketahui wilayah di Indonesia dengan persebaran kasus covid-19 yang cukup potensial dan dapat menjadi perhatian pemerintah untuk pencegahan dan penanganannya agar kasus covid-19 tidak melonjak tinggi. Pengelompokan ini bisa menjadi solusi untuk bisa mengetahui sejauh mana persebaran kasus covid-19 berbagai wilayah di Indonesia ini

## BAHAN DAN METODE

### Pengumpulan Data

Sebelum melakukan pengklasteran diperlukan tahapan analisis data yaitu pemilihan data yang nantinya akan digunakan dalam perhitungan K-Means. Sebelumnya penelitih menggunakan data Covid-19 pada tanggal 06 Juli 2021 yang diambil dari website resmi Kwala Covid-19(kawalcovid.id) sebanyak 34 data dan atribut yang digunakan adalah data kasus positif, sembuh dan meninggal. Contoh data dapat dilihat pada tabel berikut.

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

**Tabel 1.** Data Covid

| Provinsi Asal | Kasus Positif | Sembuh | Meninggal |
|---|---|---|---|
| Aceh | 174 | 72 | 2 |
| Bali | 424 | 269 | 5 |
| Banten | 457 | 268 | 6 |
| Bangka Belitung | 241 | 89 | 4 |
| Bengkulu | 201 | 70 | 1 |
| DI Yogyakarta | 1386 | 779 | 52 |
| DKI Jakarta | 9439 | 6100 | 137 |
| Jambi | 58 | 94 | 2 |
| Jawa Barat | 7239 | 3524 | 48 |
| Jawa Tengah | 4048 | 1428 | 232 |
| Jawa Timur | 1808 | 1077 | 122 |
| Kalimantan Barat | 259 | 122 | 25 |
| Kalimantan Timur | 726 | 234 | 19 |
| Kalimantan Tengah | 198 | 17 | 1 |
| Kalimantan Selatan | 55 | 69 | 3 |
| Kalimantan Utara | 207 | 69 | 0 |
| Kepulauan Riau | 586 | 22 | 4 |
| Nusa Tenggara Barat | 122 | 79 | 1 |
| Sumatera Selatan | 255 | 126 | 15 |
| Sumatera Barat | 304 | 208 | 9 |
| Sulawesi Utara | 164 | 8 | 2 |
| Sumatera Utara | 256 | 300 | 9 |
| Sulawesi Tenggara | 89 | 22 | 4 |
| Sulawesi Selatan | 249 | 71 | 0 |
| Sulawesi Tengah | 107 | 12 | 3 |
| Lampung | 266 | 41 | 6 |
| Riau | 418 | 324 | 3 |
| Maluku Utara | 117 | 40 | 4 |
| Maluku | 349 | 59 | 3 |
| Papua Barat | 258 | 51 | 1 |
| Papua | 15 | 4 | 0 |
| Sulawesi Barat | 45 | 0 | 1 |
| Nusa Tenggara Timur | 632 | 212 | 4 |
| Gorontalo | 37 | 3 | 0 |

## METODE PENELITIAN

Data mining merupakan proses penemuan pola-pola tertentu dari sebuah data atau basis data yang berukuran besar untuk memperoleh informasi yang sangat berguna. Tujuan utama dari data mining adalah menemukan dan menggali pengetahuan dari data dan informasi yang ada. Data mining merupakan salah satu bagian atau proses dari KDD (Knowledge Discovery in Databases) yaitu mengumpulkan dan menggunakan data masa lalu untuk menemukan keteraturan, pola, atau hubungan dalam suatu set data yang lebih besar. Tiga tahapan dalam KDD yaitu, preprocessing, process, dan post processing [6].

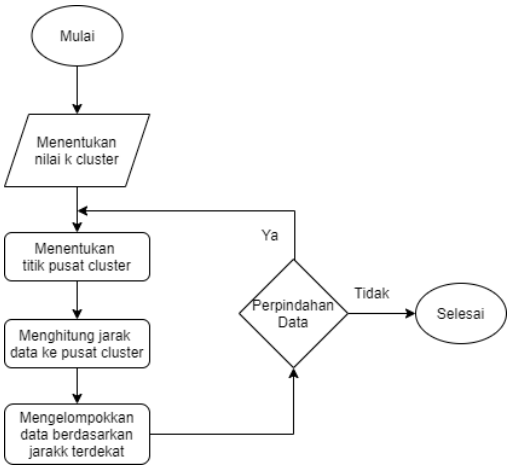
Clustering digunakan untuk mengelompokkan atau mengidentifikasi data yang memiliki karakteristik tertentu. Contoh algoritma : K-Means, K-Medoids, dan lainnya. Clustering merupakan metode data mining Unsupervised hal ini karena tidak ada satu atribut yang digunakan untuk memandu proses pembelajaran sehingga seluruh atribut diperlakukan sama..

K-Means merupakan salah satu algoritma clustering yang masuk dalam Unsupervised learning dan Non-Hierarchical atau Patitional Clustering yang digunakan untuk membagi data menjadi beberapa kelompok data dengan system partisi, data didalam satu kelompok memiliki karakteristik yang sama antara satu dengan yang lain dan memiliki karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lainnya. Dasar algoritma K-Means berikut adalah Langkah-langkah algoritma K-Means:

1) Menentukan nilai k sebagai jumlah cluster pada data set.
2) Menentukan nilai centroid (titik pusat) awal secara random
3) Menghitung jarak setiap data ke masing-masing centroid menggunakan rumus antar dua objek yaitu Euclidean Distance.
   d = √(x1-x2 )^2+(y1-y2)^2 ) .....(1)
   Dimana d adalah jarak data ke titik pusat, x1 adalah nilai titik pusat variable x, x2 adalah nilai titik pusat variable y, y2 adalah nilai data variable y.
4) Menentukan centroid baru dengan rata-rata dari data yang ada dalam cluster yang sama.
   $Ci = \frac{1}{M}\sum_{j=1}^{M} Xj$ .....(2)
   Dimana Ci adalah titik pusat ke 1, M adalah jumlah data dalam Ci, x merupakan parameter data.
5) Ulangi Langkah ke-3 hingga ke-4, jika masih ada data yang berpindah cluster.

Karakteristik dari algoritma K-Means adalah salah satu metode pengelompokkan sederhana yang dapat digunakan dengan mudah dan sangat cepat dalam proses clustering. K-Means sangat sensitif dalam pembangkitan titik pusat awal cluster (centroid awal) secara random. Pada suatu data tertentu K-Means tidak melakukan pengelompokan data dengan baik, hal ini karena hasil pengelompokannya tidak dapat memberikan pola kelompok yang mewakili karakteristik betuk alami data. Metode ini bisa mengalami masalah ketikan mengelompokkan data yang mengandung oitlier dan memungkinkan suaru cluster tidak mempunyai anggota serta sangat sulit untuk mencapai global optimum. Hasil Clustering dengan menggunakan K-Means bersifat tidak unik atau selalu berubah dan terkadang baik tetapi juga terkadang jelek [7].

Untuk lebih jelas dalam memahami alur dari algoritma K-Means bisa melihat pada gambar 1 dibawah ini




**Gambar 1**. Flowchart Algoritma K-Means

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

**Analisis Kebutuhan**

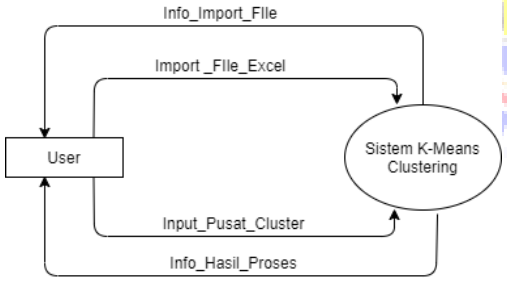
Perangkat lunak memiliki peran penting dalam penelitian inikarena hasil dari analisis data dapat diketahui dari pengolahan menggunakan perangkat lunak dalam mengetahui hasilnya. Pada penelitian ini peneliti menggunakan database MySQL, Sublime Text 3 dengan bahasa pemrograman PHP, Google Chrome,, Balsamiq Mookup 3 dan Microsoft Excel sebagai pembanding sistem yang akan dibuat sedangkan perangkat keras yang digunakan dalam penelitian ini yaitu menggunakan 1 buah laptop yang digunakan selama penelitian dengan spesifikasi laptop sebagai berikut: ASUS X450CA, Intel(R) Celeron(R) CPU 1007U @ 1.50GHz 1.50 GHz, Ram 2gb, Hardisk Storage 500gb, Windows 8.1 Pro 64bit.

**Perancangan Sistem**

Data Flow Diagram (DFD) adalah suatu teknik grafis yang digunakan untuk menggambarkan aliran informasi dan transformasi yang diterapkan saat data bergerak dari input menjadi output [8]. DFD dimulai menggambarkan system secara keseluruhan dan dilanjutkan dengan penjelasan yang lebih spesifik.

1) DFD Level 0

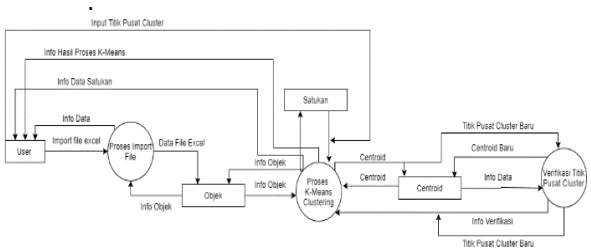
DFD Level 0 atau DFD level konteks adalah adalah gambaran dari suatu informasi melalui proses yang berjalan secara umum pada sistem aplikasi K-Means Clustering yang dirancang.Perancanagn proses sistem ini dapat dilihat pada gambar 2.




**Gambar 1**. DFD Level 0

2) DFD Level 1

DFD Level 1 adalah suatu gambaran tentang informasi jalannya alur dari sebuah sistem yang dijelaskan lebih detail atau lebih spesidik dan merupakan tahap lanjutan dari proses sebelumnya dapat dilihat pada gambar 3




**Gambar 2.** DFD Level 1

**HASIL DAN PEMBAHASAN**

**Perhitungan K-Means**

Dalam penelitian ini peneliti menentukan 3 kelompok yang berbeda yaitu: kasus positif, sembuh dan meninggal. Jika jumlah kelompok cluster sudah ditentukan Langkah selanjutnya adalah menentukan titik pusat awal. Peneliti memilih 3 data secara acak untuk dijasikan sebagai titik pusat awal seperti table dibawah ini.

**Tabel 2**. Titik Pusat Cluster Awal

| | | | |
|---|---|---|---|
| C1 (Jawa Timur) | 4048 | 1428 | 232 |
| C2 (Kalimantan Timur) | 726 | 234 | 19 |
| C3 (Maluku Utara) | 117 | 40 | 4 |

Setelah titik pisat awal dilakukan, berikutnya adalah menghitung nilai jarak data ke titik pusat menggunakan rumus persamaan 1.

d1c1 = √(174-4048)^2+(72-1428)^2+(2-232)^2
d1c1 = 4110,90160427126

d1c2 = √(174-726)^2+(72-234)^2+(2-19)^2
d1c2 = 575,531927871947

d1c3 = √(174-117)^2+(72-40)^2+(2-4)^2
d1c3 = 65,3987767469698

Perhitungan yang sama dilakukan pada keseluruhan data covid dengan titik pusat yang sama karena dapat berpengaruh pada pengujian nantinya.

Langkah berikutnya adalah menentukan titik pusat cluster baru dari setiap cluster dengan menggunakan pesamaan 2.

KPc1 = (9439+7239+4048)/3
KPc1 = 6908,666667

Sc1 = (6100+3524+1428)/3
Sc1 = 3684

Mc1 = (137+48+232)/3
Mc1 = 139

Data yang digunakan untuk mencari titik pisat baru adalah data yang dipakai untuk menghitung jarak pada data covid. Dari perhitungan yang telah dilakukan didapatkan titik pusat baru berikut.

**Tabel 3.** Titik Pusat Cluster Baru

| | | |
|---|---|---|
| 6908.666667 | 3684 | 139 |
| 804.625 | 398.125 | 26.875 |
| 175.0434783 | 70.695652 | 4.17391304 |

**Akurasi**

Perhitungan akurasi digunakan untuk menunjukkan kedekatan nilai dari hasil pengukuran dengan nilai sebenarnya. Untuk menentukan seberapa besar tingkat akurasi sebelumnya perlu diketahui nilai sebenarnya dari besaran yang diukur kemudiann akan diketahui tingkat akurasinya. Tingkat akurasi diukur berdasarkan pengelompokkan data testing yang

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

ada dengan data yang sudah ada (sebelum dilakukan perhitungan), sehingga didapatkan berikut.

**Tabel 4**. Confussion Matrix

| TP (True Positif) | FP (False Positif) |
|---|---|
| 31 | 3 |
| FN (False Negatif) | TN (True Negatif) |
| 0 | 0 |

Akurasi
= (TP+TN)/(TP+TN+FP+FN)
= (31+0)/(31+0+3+0)
= 0,91176 * 100%
= 91,176%

Dari perhitungan akurasi di dapatkan nilai TP yaitu data benar yang terklasifikasikan menjadi data benar sebesar 31 data, nilai FP benar tapi terklasifikasi salah sebesar 3 data, nilai FN atau data salah tidak diklasifikasikan salah sebesar 0 dan nilai TN atau data yang tidak di klasifikasikan sebesar 0, sehingga menghasilkan nilai akurasi sebesar 91,176%.

**Davies-Bouldin Index**

Davies-Bouldin Index adalah metode yang dipakai untuk menghitung validitas cluster pada suatu metode pengelompokan, kohesi diartikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari cluster yang diikuti, sedangkan untuk separasi berdasarkan pada jarak antar titik pusat cluster terhadap clusternya [9].

Semakin kecil nilai Davies Bouldin Index (DBI) yang dihasilkan (non-negatif >= 0), maka semakin baik cluster yang diperoleh dari pengelompokan menggunakan algoritma clustering [10]. Rumus persamaan perhitungannya sebagai berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max i \neq j \, (Ri,j) \qquad \text{.....(3)}$$

Dari perhitungan diatas diperoleh nilai Davies-Bouldin Index (DBI) senilai 0,493371469. Nilai tersebut mendekati nol dan bernilai tidak negetif sehingga dapat disimpulkan juga bahwa cluster sudah cukup baik.

**KESIMPULAN**

Dari penelitian yang telah dilakukan maka penulis dapat mengambil kesimpulan sebagai berikut:

1) Pada penelitian ini menerapkan metode algoritma K-Means Clustering menggunakan dua aplikasi yaitu Microsoft Excel 2010 dan Sistem Aplikasi K-Means Clustering yang telah dibuat. Dengan evaluasi cluster menggunakan Davies-Bouldin Index diperoleh nilai 0,493371469 sehingga dapat disimpulkan hasil cluster cukup baik dan akurasi 91,176%.
2) Pada penelitian ini diperoleh pusat cluster pertama yang merupakan daerah dengan tingkat penyebaran kasus Covid-19 tersebar pada provinsi DKI Jakarta dan Jawa Barat, cluster kedua berada pada provinsi DI Yogayakarta, Jawa Tengah dan Jawa Timur, sedangkan cluster ketiga yaitu provinsi lainnya di Indonesia yang tidak masuk termasuk dalam cluster pertama dan kedua.

**DAFTAR PUSTAKA**

[1] Rizkiana Prima R., Y. A. 2020. Analisis Cluster Virus Corona (COVID-19) di Indoensia pada 2 Maret 2020 -12 April 2020 dengan Metode K-Means Clustering.https://www.researchgate.net/publication/342697385.
[2] Wiyli Yustanti, N. R. 2020. Klastering Wilayah Kota/Kabupaten Berdasarkan Data Persebaran Covid-19 di Propinsi Jawa Timur dengan Meode K-Means. Journal Information Engineering and Education Technology Vol.4 No.1, ISSN 2549-869X.
[3] Sukma Sindi, W. R. 2020. Analisis Algoritma K-Medoids Clustering dalam Pengelompokkan Penyebaran Covid-19 di Indonesia. Jurnal Teknologi Informasi Vol.4 No.1, ISSN 2580-7927.
[4] Nayuni Dwitri, J. A. 2020. Penerapan Algoritma K-Means dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 di Indonesia. Jurnal Teknologi Indormasi Vol.1 No.1, ISSN 2580-7927.
[5] Idah Wahidah, M. A. 2020. Pandemik Covid-19: Analisis Perencanaan Pemerintah dan Masyarakat dalam Berbagai Upaya Pencegahan. Jurnal Manajemen dan Organisasi, 179-188.
[6] Prasetyowati, E. 2017. DATA MINING Pengelompokan Data untuk Informasi dan Evaluasi. Pamekasan: Duta Media Publishing.
[7] Anjar Wanto, M. N. 2020. Data Mining: Algoritmadan Implementasi. Medan: Yayasan Kita Menulis.
[8] Roger S. Pressman. 2002. Rekayasa Perangkat Lunak Pendekatan Praktisi (Buku Satu). Yogyakarta: Penerbit ANDI.
[9] Wani, M.A & Riyaz, R. 2017. A novel point density based validity index for clustering gene expression datasets. International Jounal of Data Mining and Bioinformatics 17(1), 66-84.
[10] Kalita, A.B. (2016). Counting clsuters in twitter posts. Proceedings of the 2nd International Conference on Information Technology for Competitive, (pp. 85).