

Komprasi Perbandingan Algoritma K-Means dan K-Medoids Dalam Clustering Penyebaran Kasus Covid 19

Ladyka Febby Olivia¹, Dwana Abdi Juliantho², Billy Hendrik³

^{1,2}, Fakultas Ilmu Komputer, Universitas Putra Indonesia "YPTK" Padang, Sumatera Barat

²Dosen Universitas Putra Indonesia "YPTK", Padang, Sumatera Barat

¹ladyka.febby@gmail.com, ²abdidwana@gmail.com, ³billy_hendrik@upiyptk.ac.id

ARTICLE INFO

Submit : 12-08-2023
Review : 14-08-2023
Accepted : 25-08-2023
Published : 28-08-2023

Keyword :

*Covid -19,
DBI,
RapidMinner,
K-Means,
K-Medoids*

ABSTRACT

Covid 19 is a rapid spread virus that has a negative impact on the province of XYZ. With the Indonesian government policy, four zone policies such as the red, green, yellow, orange zone, which each zone has a different meaning. The algorithm used in this study is K-Means and K-Medoids. Algorithm-Means group data by dividing them into several clusters based on the same characteristics. Whereas the K-Medoids algorithm selects real objects to represent the cluster. In this study, the two algorithms were compared using one dataset. The comparison is done by looking at the value of Davies-Bouldin Index (DBI) on the rapidminer. The best result of K-Means DBI is 0.078 while K-Medoids gives a value of 0.250 which is divided into 2 clusters

1. Pendahuluan

Pada 31 Desember 2019, WHO China Country Office melaporkan kasus penyebaran virus dikota Wuhan, Provinsi, China. Dan pada itu WHO memberikan nama virus tersebut Severe Acute Resoiratory Syndrome Coronavirus-2 (SARS-Co V-2) dan resmi menetapkan [enyakit tersebut dengan nama Novel Coronavirus pada manusai dengan sebutan Corona virus Diseases 2019 atau lebih dikenal sebutkan Covid 19. dimana jika terserang penyakit ini akan menimbulkan gangguan pernapasan seperti flu dan gangguan berat lainnya[1].

Pada tanggal 2 Maret 2020, Indonesia untuk pertama kalinya mengonfirmasi kasus COVID-19. Pandemi COVID-19 yang terjadi ini membawa dampak yang serius pada tatanan kesehatan, perekonomian, dan sosial di indonesia[2].

Adapun algoritma yang digunakan untuk melakukan pengelompokan ialah *K-Means* beserta *K-Medoids*. Algoritma *K-Means* adalah metode data mining yang termasuk kata lain tanpa supervisi ketika melakukan proses pemodelan, algoritma ini juga merupakan algoritma dengan sistem partisi untuk melakukan pengelompokan data. Pada proses pengelompokan data, terdapat dua jenis metode, yakni Hierarchical dan Non-Hierarchical, untuk algoritma *K-Means* tergolong dalam metode Non-Hierarchical atau dapat juga disebut dengan istilah Partitional Clustering. Algoritma *k-Means* melakukan pembagian data dengan membaginya ke dalam kelompok atau cluster, yang mana data di dalam satu cluster memiliki ciri-ciri yang sama pada tiap datanya

JISED
Journal of Information System and Education Development
namun memiliki ciri-ciri yang berbeda terhadap data yang terdapat pada cluster lainnya[3].

Algoritma kedua yang diaplikasikan yaitu *K-Medoids*, algoritma ini dikenal juga dengan sebutan Algoritma Partitioning Around Medoid yang disingkat PAM. Pada algoritma *K-Medoids* pengelompokan partisi memakai objek nyata yang dipilih untuk mewakili klaster[4]. Bertujuan untuk menghasilkan partisi yang tidak terlalu rentan terhubung karena nilai-nilai pada dataset yang ekstrim[5].

Penelitian tersebut bertujuan mengelompokkan Provinsi yang rawan tindak kriminalitas agar dapat menentukan Provinsi mana saja yang membutuhkan ditingkatkan pengawasannya. Hasil penelitian yang dapat ditarik yaitu melalui perhitungan *K-Means*, didapatkan kelompok tindak kejahatan tinggi (cluster 1) terdapat 6 anggota, dan untuk kelompok tindak kejahatan rendah (cluster 2) ada 28 anggota. Sedangkan untuk hasil *K-Medoids* didapatkan untuk tindak kejahatantinggi (cluster 1) memiliki 7 anggota dan 27 anggota tergabung dalam kelompok tindak kejahatan rendah (cluster 2)[6].

Penelitian dengan kasus serupa juga dilakukan oleh Bu'ulolo dan Purba (2021) dengan algoritma yang berbeda untuk membentuk cluster zona penyebaran Covid-19 di Provinsi Sumatera Utara dengan algoritma *K-Medoids* menggunakan variabel positif, ODP, PDP dan meninggal. Hasil yang didapatkan dari penelitian ini yaitu didapatkannya 3 cluster zona penyebaran[7].

Pada penelitian ini penulis ingin menerapkan data minung pada data kasus Covid-19 di Provinsi XYZ dengan

menggunakan metode kompresi K-Mean dan K-Medoids. data mining adalah proses menemukan pola dan pengrhayna yang menarik dari suatu data dalam jumlah yang besar.

2. Metodologi Penelitian

Gambar 1 memvisualisasi prosedur atau tahapan -tahapan yang dilkaukan dalam pengelompokan.



Gambar 1. Prosedur Penelitian

Tahapan identifikasi masalah adalah suatu proses untuk merumuskan masalah dalam penelitian. Agar dalam proses analisa data terstruktur, sehingga dapat memahami lebih rinci mengenai ruang maslah penelitian dibutuhkanya studi literatur, mengumpulkan serta mempelajari berbagi referensi yang dijadikan sebagai bahan penunjang penelitian yang diangkat. Dalam hal ini literatur yang dimaksudkan ialah beruoa jurnal-jurnal penelitian maupun buku dalam penelitian yang diangkat yang mana membahas Algoritma K-Means Dan K-Medoids.

2.1 Algoritma K-Means

Algoritma K-Means memecah data berdasarkan jarak tiap data terhadap cluster yang sudah ditentukan. K-Means bertumpu pada fungsi untuk menghitung data yang memiliki karakteristik yang sama. Berikut ini langkah-langkah untuk pengelompokan data dengan K-Means[8].

1. Menetapkan jumlah klaster.
2. Inisialisasi awal dan centroid (pusat klaster) ditentukan secara acak.
3. Menaksir jarak dari tiap centroid dengan menggunakan formula Euclidean Distance.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

4. Menghitung centroid baru dengan kedudukan baru pula dengan cara mencari rata-rata objek pada cluster

5. Mengulangi langkah sebelumnya sampai hasil pengelompokan tidak mengalami perubahan.

2.2 Algoritma K-Medoids

K-Medoids menggunakan metode partisi clustering dalam melakukan klasterisasi sejumlah n objek menjadi suatu kelompok [11]. Tahapan proses algoritma K-Medoids adalah [9]:

1. Menetapkan jumlah klaster.
2. Menentukan centroid dengan random sebanyak jumlah klaster.
3. Mengukur jarak dari setiap data dengan formula berikut:

$$d_1 = \sqrt{(x_1 - x_{co})^2 + (y_1 - y_{co})^2} \quad (2)$$

Adapun nilai dari x_1 mewakili luas areal dan nilai dari y_1 mewakili jumlah produksi berdasarkan dataset pada tabel 1. Sehingga akan dicari satu per satu berdasarkan data yang ada untuk mengetahui jarak terdekatnya. Dalam hal ini, nilai pada dataset untuk x_1 dan y_1 akan dikurangkan dengan data yang menjadi centroid kemudian dikuadratkan lalu dijumlah, setelah itu diakarkan.

4. Lakukan repetisi medoids dengan mengukur jarak dengan memakai metode taksir jarak, kepada semua data yang bertindak sebagai data proses.
5. Menetapkan objek yang tidak representatif dengan random

2.3 Davies- Boudline Index (DBI)

D. L. Davies dan D. W. Bouldin pertama kali mempublikasikan Davies-Bouldin Index pada tahun 1979. Davies Bouldin berguna untuk evaluasi cluster yang didasarkan pada nilai kohesi dan separasi. Formula yang digunakan untuk mengetahui nilai kohesi menggunakan Sum of square within cluster (SSW) pada cluster ke-i

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \quad (3)$$

Pada persamaan SSW adalah jumlah data yang ada pada cluster serta cimerupakan centroid dari cluster. Kemudian untuk nilai separasi dihitung dengan Sum of square between cluster (SSB) yang dirumuskan:

$$SSB_{i,j} = d(c_i, c_j) \quad (4)$$

Jika nilai kohesi dan separasi telah didapatkan selanjutnya mencari pengukuran rasio (R_{ij}), dengan rumus:

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (5)$$

Lalu dengan nilai rasio tersebut digunakan untuk mencari nilai DBI dengan formula:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} (R_{i,j}) \quad (6)$$

3. Hasil dan Pembahasan

Dari 13 pengujian sampel data yang dioleh menggunakan K=4. dimana jumlah K menunjukkan penyebaran daerah covid 19 pada daerah XYZ. yang mana daerah tersebut mempunyai zona merah, orange, kuning dan hijau. Setelah data diuji sebanyak 13 dan diuji dengan RapidMiner, dan didapatkan hasil seperti dibawah ini.

Tabel 1. Hasil pengujian algoritma K=4

Uji Ke-	DBI K-Means	DBI K-Medoids
1	0.474	0.412
2	0.461	1.171
3	0.595	1.074
4	0.482	0.388
5	2.475	7.871
6	2.206	12.346
7	0.461	1.646
8	1.351	3461
9	0.474	1.172
10	4.085	15.692
11	0.461	1.142
12	1.795	1.959
13	0.474	0.930

Dari tabel diatas memiliki perbedaan hasil DBI dari algoritma K-Means dan K-Medoids. Dan ada konsep Analisa Perbandingan Algoritma yang dimana untuk melihat hasil yang signifikan sebesar 0,001 dan 0,000 lebih kecil dari 0,05, maka dapat dinyatakan data tidak terdistribusi normal. *MannWhitney* ini dilihat pada kolom Asymp Sig (2 tailed) atau nilai signifikan untuk uji dua sisi. Apabila nilai sig > 0.05 maka H0 diterima dan jika sig < 0.05, H0 ditolak.

Setelah diuji kembali K=2 sampai K=8 maka didapatkan hasil yang lebih nyata yang lebih kecil pada *Algoritma K-mean* dengan K=2 sebesar 0.078. sedangkan *Algoritma K-Medoids* K optimal adalah =3 dengan nilai DBI sebesar 0.250.

4. Kesimpulan

1. Dari hasil clustering yang dilakukan, tidak terdapat perbedaan yang signifikan antar algoritma *K-Means* dan *K-Medoids*. Dimana hasil DBI daerah XYZ yaitu 0,250.
2. Berdasarkan hasil pengujian mencari Nilai (*Davies Bouldin Index*) DBI menggunakan bantuan tools RapidMiner, tetapi hasilnya berbeda.

Saran

Untuk peneliti selanjutnya yaitu menambahkan algoritma clustering yang lain dan membandingkan beberapa algoritma untuk memperoleh hasil yang terbaik.

Reference

- Sekretaris Jenderal Kementerian Dalam Negeri Direktur Jenderal Bina Administrasi Kewilayahan Direktur Jenderal Politik dan Pemerintahan Umum Direktur Jenderal Bina Pembangunan Daerah Direktur Jenderal Bina Keuangan Daerah Penyusun Safrizal, P. Z., Danang Insita Putra, Ms., Sofyan, S., Bimo MPH Kontributor Edy Suharmanto, Mc., Hidayat, W., Uswantini, A., Nursetyo, E., Laksito, P., Wiratama, Y., Meirandha, G., Zulkarnain, E., Damara, R., Sutanti, N., Anindita, K., Lumban Gaol, H., Zulkarnain, R., & Bagus Suryowibowo, dan. (n.d.). *TIM PENYUSUN Pembina MENTERI DALAM NEGERI*.
- B'ulolo, E., & Purba, B. (n.d.). *Algoritma Clustering untuk membentuk Cluster Zona Penyebaran Covid-19*. <https://doi.org/10.31849/digitalzone.v12i1.6572ICCS>
- Wayan Wardani, N. (n.d.). *Analisis Pesebaran Penularan Virus Corona Di Provinsi Jawa Tengah Menggunakan Metode K-Means Clustering*. <http://jurnal.mdp.ac.id>
- Nabil Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, 2(2), 100. <http://jim.teknokrat.ac.id/index.php/JTSI>
- S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. I. R.H.Zer, and D. Hartama, "Analisis Algoritma KMedoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia," *J. Teknol. Inf.*, vol. 4, no. 1, pp. 166–173, 2020, doi: 10.36294/jurti.v4i1.1296
- E. G. Sihombing, "Klasifikasi Data Mining Pada Rumah Tangga Menurut Provinsi Dan Status Kepemilikan Rumah Kontrak/Sewa Menggunakan KMeans Clustering Method | Sihombing | CESS (Journal of Computer Engineering, System and Science)," *Jul. 02, 2017*. <https://jurnal.unimed.ac.id/2012/index.php/cess/article/view/6347/5933> (accessed Jul. 27, 2022).
- D. F. Pramesti, M. T. Furqon, and C. Dewi, "Tampilan Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2017. <https://jptiik.ub.ac.id/index.php/j-ptiik/article/view/204/101>
- Abidin, N. A. S. Z., Avila, R. D., Hermatyar, A., & Rismayani, R. (2022). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Daerah Produksi Kakao. *Jurnal Teknik Informatika Dan Sistem Informasi*, 8(2). <https://doi.org/10.28932/jutisi.v8i2.489>