

Penerapan Metode Cosine Similarity Untuk Meningkatkan Kinerja K-Means Pada Pengelompokan Wilayah Penanganan Covid Di Dki Jakarta

Implementation Of Cosine Similarity Method To Increase K-Means Performance On Area Grouping Of Covid Handling In Dki Jakarta

Almanda Nosra¹⁾, Deni Arifianto^{2)*}, Miftahur Rahman³⁾

¹Mahasiswa Fakultas Teknik, Universitas Muhammadiyah Jember

Email : almandanosra12@gmail.com

²Dosen fakultas Teknik , Universitas Muhammadiyah Jember* Koresponden Author

Email : deniafrianto@unmuhjember.co.id

³Dosen Fakultas Teknik, Universitas Muhammadiyah Jember

Email : miftahurrahman@unmuhjember.co.id

Abstrak

Fenomena Covid-19 telah menggemparkan dunia, Indonesia adalah salah satu negara dimana masyarakatnya terdampak dari virus tersebut. Pada penelitian ini dilakukan klusterisasi epidemi virus Covid-19 di DKI Jakarta. Kota tersebut dipilih berdasarkan angka kasus tertinggi di Indonesia. Alasan dilakukannya klusterisasi ini berkaitan dengan mengelompokkan kasus persebaran covid di daerah-daerah DKI Jakarta dimana nantinya akan dilakukan untuk menentukan penanganan Covid-19. Menerapkan teknik data mining. Pengelompokan didasarkan pada nomor parameter dirawat, sembuh, meninggal dan isolasi mandiri. Metode K-Means dan metode Cosine Similarity, dan diuji dengan metode DBI (Davies Bouldin Index) dengan menghitung tingkat perhitungan DBI dengan menggunakan metode K-Means tanpa cosine dan tingkat perhitungan DBI dengan menggunakan metode K-Means cosine. Penerapan teknologi data mining. Pengelompokan didasarkan pada nomor parameter. Pengklusteran dilakukan berdasarkan penyebaran kasus terbanyak di provinsi DKI Jakarta. Hasil perhitungannya adalah Tingkat perhitungan nilai DBI yang paling baik dengan menggunakan metode K-means cosine Similarity, karena nilai yang diperoleh rendah, yaitu diperoleh nilai DBI (Davies Bouldin Index) terkecil pada cluster 9 yaitu sebesar -5.527. Sedangkan nilai DBI terbesar pada 2 cluster dengan nilai -2.282.

Kata kunci : Covid 19 DKI Jakarta, Data Mining, K-means cosine similarity

Abstract

Covid-19 phenomenon that has shocked the world, Indonesia is one of the countries where the people are affected by the virus. In this study, a clustering of the spread of the Covid-19 virus was carried out in DKI Jakarta, the city was selected based on the highest number of cases in Indonesia. The reason for this clustering is related to grouping cases of the spread of covid in DKI Jakarta areas which will later be carried out to determine the handling of Covid-19. By applying data mining methods. The grouping was carried out based on the number of parameters treated, recovered, died and self-isolation. The K-Means method and the Cosine Similarity method, and tested using the DBI (Davies Bouldin Index) method by calculating the DBI calculation rate using the K-Means method without cosine and the DBI calculation level using the K-Means cosine method. Produce a prototype for grouping data on the distribution of patients infected with Covid-19. Clustering is carried out based on the distribution of the most cases in DKI Jakarta province. The result of the calculation is that the best level of calculation of the DBI (Davies Bouldin Index) value using the K-means cosine Similarity method, because the value obtained is low, namely the smallest DBI (Davies Bouldin Index) value in cluster 10 is -3.469. While the largest DBI value is in 2 clusters with a value of -2.282.

Keywords : Covid-19 DKI, Data Mining, K-means cosine similarity

1. PENDAHULUAN

Pada penelitian ini dilakukan klusterisasi epidemi virus Covid-19 di DKI Jakarta., kota tersebut dipilih berdasarkan angka kasus tertinggi di Indonesia. Alasan dilakukannya klusterisasi ini berkaitan dengan mengelompokkan kasus persebaran covid di daerah-daerah DKI Jakarta dimana nantinya akan dilakukan untuk menentukan penanganan Covid-19. Menerapkan teknik data mining. Pengelompokan didasarkan pada nomor parameter dirawat, sembuh, meninggal dan isolasi mandiri. Metode K-Means dan metode *Cosine Similarity*, ini menghasilkan *prototipe* pengelompokan data persebaran pasien terinfeksi Covid19. Pengklasiran dilakukan berdasarkan penyebaran kasus terbanyak di provinsi DKI Jakarta.

Clustering merupakan salah satu metode data mining, yaitu teknik mengelompokkan data, mengamati atau memperhatikan dan membentuk kelas-kelas objek yang memiliki kemiripan. Salah satu teknik clustering yang paling terkenal adalah algoritma kmeans karena kmeans memiliki algoritma yang sederhana dan efisien.. Untuk membuat Kmeans mudah dipelajari (Gustientiedina, Adiyaa dan Desneliita, 2019). Penggunaanya sangat umum, menggunakan prinsip-prinsip sederhana yang dapat dijelaskan dalam istilah non-statistik. Namun di samping kemudahan, terdapat kekurangan, yaitu sebelum algoritma dijalankan, titik K di inisialisasikan acak sehingga pengelompokan data yang diperoleh tidak optimal. Jika terjebak dalam kasus yang dikenal sebagai kutukan dimensi. Jika ingin mencari jarak antar titik Dari 2D masih mudah, tetapi dalam 20 dimensi sangat sulit untuk menemukan jarak.. Jika ada beberapa titik sampel data yang ada, Anda dapat dengan mudah melakukan perhitungan dan juga mencari jarak dari titik terdekat ke k titik yang diinisialisasi secara acak. Untuk tahapan akhir yaitu metode validasi untuk menguji sebuah cluster yang paling baik. Pada penelitian ini menggunakan metode Indeks Davies Bouldin diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Indeks Davies Bouldin (DBI) adalah metrik untuk mengevaluasi hasil dari algoritma clustering.

Pada tugas akhir ini berkaitan dengan mengelompokkan daerah penanganan Covid-19 di DKI Jakarta guna untuk menentukan penanganan Covid-19 terbaik. metode yang digunakan yaitu Pengelompokan penyebaran virus corona di DKI Jakarta menggunakan metode ini *K-Means Clustering* dan di sempurnakan dengan *cosine similiary*, dan diuji dengan metode DBI. Untuk itu hasil perhitungan nanti akan dilihat berapa tingkat perhitungan DBI dengan menggunakan metode K-Means dan metode *K-Means cosine similarity*. Hasil penelitian ini diharapkan dapat membantu pemerintah DKI Jakarta mengambil keputusan strategis untuk menekan penyebaran virus tersebut Covid-19 di DKI Jakarta.

2. TINJAUAN PUSTAKA

A. Covid Dki Jakarta

Pada penelitian ini dilakukan klusterisasi epidemi virus Covid-19 di DKI Jakarta, dimana data yang diperoleh adalah bulan oktober, november, dan desember tahun 2020. Jumlah kasus covid mengalami angka kenaikan tiap bulannya, hal ini dibuktikan dengan adanya catatan kasus covid berdasarkan parameter jumlah positif pada bulan oktober sejumlah 105597, meningkat ke november terhitung 136861 hingga melaju pesat pada bulan desember 183735, di tahun 2020.

B. Data Mining

Data mining adalah proses yang Ekstrak dan identifikasi informasi yang berguna dan pengetahuan terkait dari database besar menggunakan metode statistik, matematika, kecerdasan buatan, dan pembelajaran mesin. (Turban et al., 2005).

C. K-Means Clustering

Salah satu teknik clustering yang paling terkenal adalah algoritma kmeans karena kmeans memiliki algoritma yang sederhana dan efisien.. Untuk membuat Kmeans mudah dipelajari (Gustientiedina, Adiya dan Desnelita, 2019).

Langkah-langkah Algoritma K-Means dapat Proses dasar algoritma KMeans (Wardhani, 2016):

1. Tentukan cluster sebagai jumlah cluster yang Anda inginkan bentuk. Tentukan pusat cluster.

2. Hitung jarak cluster untuk masing-masing data ke pusat cluster menggunakan persamaan Euclidean.
3. Kelompokkan data pada cluster dengan jumlah jarak terpendek menggunakan persamaan.
4. Hitung titik pusat cluster baru menggunakan persamaan.
5. Ulangi langkah 2-5 sehingga tidak ada lagi data yang ditransfer ke cluster lainnya.

D. Cosine Similarity

Metode Cosine Similarity adalah metode yang digunakan untuk menghitung kemiripan (degree of similarity) antara dua objek. Secara umum, perhitungan metode ini didasarkan pada ukuran kesamaan ruang vektor. Metode cosinus similarity ini menghitung kemiripan antara dua objek (misalnya D1 dan D2) yang direpresentasikan sebagai dua vektor dengan menggunakan kata kunci kata terdapat dari dokumen sebagai ukuran (G. A. Pradnyana dan N. A. Sanjaya, 2012):

$$\cos(\theta_{ij}) = \frac{\sum_k (d_{ik}d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}}$$

$\sum_k (d_i * d_j)$ = Jumlah bobot dokumen pertama dikalikan dokumen kedua

$\sqrt{\sum_k d_i^2}$ = Akar jumlah dari bobot dokumen pertama yang sudah di kuadratkan

$L \sqrt{\sum_k d_i^2}$ = Akar jumlah dari bobot dokumen kedua yang sudah di kuadratkan

E. Davies-Bouldin Index (DBI)

Davies Bouldin Index diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Davies Bouldin Index (DBI) adalah metode validasi untuk pengujian cluster terbaik. Prosedur untuk menghitung Davies Bouldin Index (DBI) adalah sebagai

Rumus DBI

$$DB = \frac{1}{n} \sum_i^n \frac{n}{i} = 1, i = i \text{MAX} \left(\frac{a_i + a_j}{d(c_i, c_j)} \right)$$

Di mana :

- DB : Davies Bouldin
- n : jumlah cluster
- σ_i : Rata-rata jarak dari data ke-i
- σ_j : rata-rata jarak dari data dengan titik pusat data cluster ke-j
- c_i : titik pusat data cluster
- c_j : titik pusat data cluster
- $d(c_i, c_j)$: jarak antara centroid

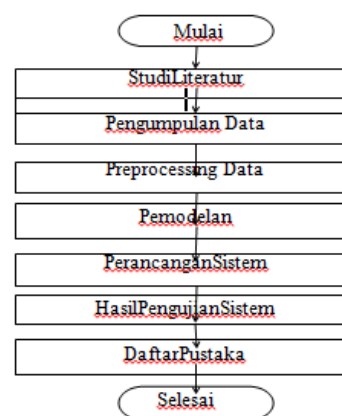
F. Rapid Miner

Rapid Miner adalah solusi penambahan data, penambahan teks, dan analitik prediktif. Rapid Miner adalah software atau standalone (open source) software untuk analisis data sebagai data miner yang dapat diintegrasikan ke dalam produknya sendiri. Menggunakan berbagai teknik deskriptif dan prediktif, Rapid Miner memberikan informasi kepada pengguna untuk membantu mereka membuat keputusan terbaik.

3. METODE PENELITIAN

A. TAHAP PENELITIAN

Tahapan yang dilakukan pada penelitian ini ditunjukkan pada



Gambar 1 Flowchart alur penelitian
 Sumber : penelitian sendiri

B. STUDI LITERATUR

Penelitian bibliografi digunakan untuk mencari referensi teoritis terkait kasus atau masalah yang ditemukan. Periksa literatur dalam penelitian ini dengan mencari review dari penelitian sebelumnya yang dapat digunakan sebagai referensi atau referensi untuk mendukung kemudian dan menyempurnakan pemahasan saat ini.. Oleh karena itu, penulis harus meneliti sejumlah dokumen yang digunakan kemudian memilih literature sehingga dapat ditentukan bahan apa yang akan digunakan dalam penelitian tersebut.

C. ANALISA KEBUTUHAN

Pada tahap analisa kebutuhan dilakukan sesuai yang dibutuhkan sistem yang dibangun dapat melakukan diagnosis penyakit Ibu hamil. Dan untuk mempermudah masyarakat untuk mengetahui penyakit apa yang mereka derita sejak dini dan dimaksudkan untuk lebih cepat cara menanggulangnya.

D. PREPROCESSING DATA

Prapemrosesan data adalah operasi pemrosesan data mentah yang bertujuan untuk menghasilkan kumpulan data akhir. Dalam penelitian ini data kasus Covid19 di DKI Jakarta dikumpulkan dengan menggunakan contoh 1 data per hari dan dibersihkan terlebih dahulu dari atribut yang tidak terpakai selama implementasi.berfungsi Kemudian data akan dinormalisasi menjadi rentang 01 menggunakan ZScore. Dalam proses penghitungan data yang dinormalisasi, alat yang digunakan adalah Microsoft Excel.

E. MODELING (PEMODELAN)

Pada tahap pemodelan akan diimplementasikan algoritma clustering kmeans. Metode KMeans Cosine Similarity digunakan untuk data mining untuk mengelompokkan data menjadi cluster atau kelompok berdasarkan kesamaan variabel data atau atribut.

F. CLUSTERING K-MEANS DENGAN COSINE SIMILARITY.

Berikut langkah-langkah penyelesaiannya

1. Proses awal mengukur kesamaan antara data satu dengan data lain memakai *cosine*

similarity rumus *s cosine* dapat dilihat pada persamaan.

$$\text{sim}(x_a, x_b) = \cos \theta \frac{x_a \cdot x_b}{\|x_a\| \|x_b\|}$$

Dapat dilihat dibawah ini perhitungan *cosine similarity* pada data diatas :

$$\text{sim}(x_a, x_b) = \cos \theta \frac{x_a \cdot x_b}{\|x_a\| \|x_b\|}$$

Dapat dilihat dibawah ini perhitungan *cosine similarity* pada data diatas :

$$\text{sim}(\text{cakung, cakung}) = \frac{(0.4780 \cdot 0.4780) + (-0.0582 \cdot -0.0582) + (-0.1384 \cdot -0.1384) + (0.7597 \cdot 0.7597)}{\sqrt{0.4780^2 + -0.0582^2 + -0.1384^2 + 0.7597^2} \times \sqrt{0.4780^2 + -0.0582^2 + -0.1384^2 + 0.7597^2}}$$

$$\cos \theta = \frac{0.8281}{0.9180 \times 0.9180}$$

$$\cos \theta = \frac{0.8281}{0.8281}$$

$$\cos \theta = 1$$

$$\text{sim}(\text{cakung, cempakutih}) = \frac{0}{\sqrt{0.4780^2 + (-0.0582^2) + (-0.1384^2) + 0.7597^2} \times \sqrt{0.8105^2 + (-0.1160^2) + (-0.2969^2) + (-0.4799^2)}}$$

$$\cos \theta = \frac{0}{0.9281 \times 0.6542}$$

$$\cos \theta = \frac{0}{0.6074}$$

$$\cos \theta = 0$$

2. Tentukan jumlah cluster, jumlah cluster adalah jumlah grup yang akan dibuat. Pada penelitian ini jumlah cluster yang akan digunakan adalah 2.
3. Tentukan centroid awal, centroid awal diperoleh dengan mencari yang tertinggi, terendah dan median. Pusat awal adalah pusat cluster pertama
4. Hitung jarak setiap data ke pusat cluster. Ini adalah perhitungan menggunakan persamaan spasial jarak Euclidean:
5. Clustering dan data cluster, setelah menghitung data jarak centroid, langkah selanjutnya adalah clustering data.
6. Setelah semua data ditempatkan di cluster terdekat, hitung kembali pusat cluster baru berdasarkan rata-rata keanggotaan cluster untuk menentukan pusat/titik tengah baru.
7. Setelah mendapatkan titik pusat baru setiap cluster, hitung ulang data dengan pusat cluster baru, ulangi sampai mendapatkan sampel akhir yang tidak bergerak.

G. EVALUASI

Pada tahap selanjutnya yaitu evaluasi, hasil clustering yang diperoleh dari jumlah cluster k = 3 dievaluasi menggunakan Davies Bouldin Index (DBI). Metode validasi termasuk uji cluster terbaik.

4. PEMBAHASAN DAN HASIL

A. Rapid Miner K-Means Dengan Cosine Similarity

Atribut-atribut tersebut akan diolah dengan menggunakan metode *K-Means Cosine Similarity* dan dilakukan juga uji untuk menentukan berapa tingkat perhitungan nilai DBI menggunakan *Rapid Miner*. Pengelompokan ini dilakukan berdasarkan Kecamatan.

B. Clustering Algoritma K-Means Cosine Similarity

Proses *clustering* algoritma *K-Means Cosine Similarity* untuk mengelompokkan data Kecamatan di DKI Jakarta dapat dihitung menggunakan *tools Rapid Miner*.

C. Implementasi RapidMiner

Langkah-langkah menggunakan *tools RapidMiner* untuk menghitung dan mengelompokkan data kecamatan sebagai berikut :

- Buka aplikasi *RapidMiner*
- Pada bagian operator *search “Read Excel”* lalu seret ke panel proses
- *Import* data yang akan dihitung di bagian *parameters Search “Normalize”* dan seret ke pane
- proses
- *Search “Data To Similarity”* dan seret ke panel proses
- *Search “K-Means”* dan seret ke panel proses
- Atur *cluster* yang diinginkan di bagian parameter
- *Search “Cluster Distance Performance”* dan seret ke panel proses
- Ganti metode validasi menjadi *Davies Bouldin Index* di bagian *parameters*
- Lalu sambungkan tiap-tiap operator dan *Run*

D. Hasil Perhitungan Nilai DBI Menggunakan RapidMiner

Dalam penelitian ini menggunakan 10 cluster. Cluster yang terbaik ditentukan

berdasarkan perhitungan *RapidMiner* yang menggunakan metode validasi *Davies Bouldin Index (DBI)*. Cluster terbaik yang diperoleh dari hasil perhitungan cluster memiliki nilai DBI terkecil untuk 10 cluster.

Jumlah Cluster	Nilai DBI K-Means	Nilai DBI K-Means Cosine Similarity
2	-0,054	-1.284
3	-0.113	-1.280
4	-0.546	-1.483
5	-0.746	-1.362
6	-0.861	-1.332
7	-0.868	-1.293
8	-0.769	-1.261
9	-0.926	-1.243
10	-0.957	-1.239

Tabel 1 Nilai DBI K-Means Dan K-Means Cosine

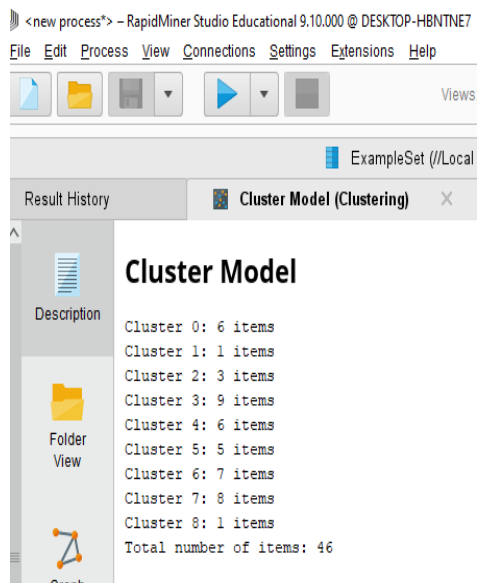
Sumber : Hasil Perhitungan

Dapat dilihat tingkat perhitungan nilai DBI (*Davies Bouldin Index*) yang paling baik adalah dengan menggunakan metode *K-means cosine Similarity* dikarenakan nilai yang diperoleh rendah, semakin rendah nilai DBI (*Davies Bouldin Index*) yang diperoleh maka akan semakin baik metode yang digunakan.

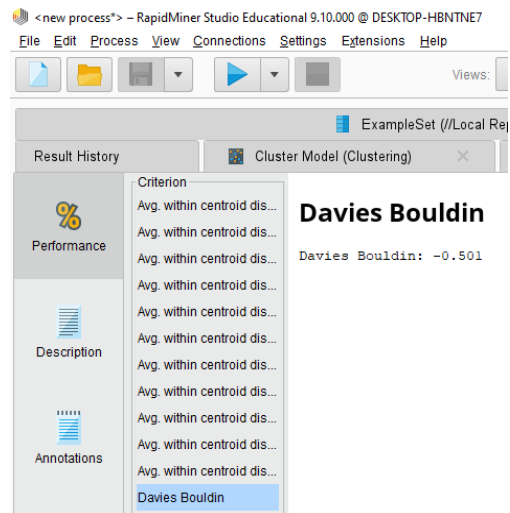
E. Hasil Pengujian Data 9 Cluster

Tingkat perhitungan DBI (*Davies Bouldin Index*) dengan menggunakan metode *K-Means cosine similarity*, diperoleh nilai DBI (*Davies Bouldin Index*) terkecil pada cluster 4 yaitu sebesar -1.483. Sedangkan nilai DBI terbesar pada 10 cluster dengan nilai -1.239. Anggota tiap-tiap cluster terdiri dari:

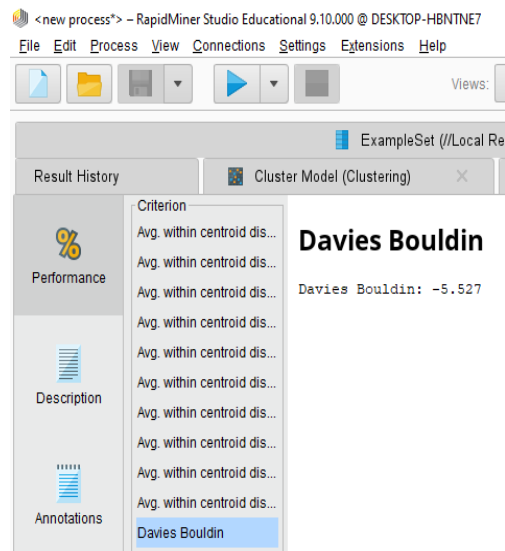
- Cluster 1 : 6 Kecamatan
- Cluster 2 : 24 Kecamatan
- Cluster 3 : 9 Kecamatan
- Cluster 4 : 7 Kecamatan
- Total Kecamatan : 46



Gambar 2 4Cluster Model *K-Means Cosine Similarity*
Sumber : Hasil Perhitungan



Gambar 3 *Davies Bouldin Index K-Means 10 Cluster*
Sumber : Hasil Perhitungan



Gambar 4 *Davies Bouldin Index K-Means Cosine Similarity 9 Cluster*
Sumber : Hasil Perhitungan

Pada Gambar 4.4 terdapat hasil dari perhitungan yang menggunakan Metode *K-Means* tanpa *Cosine similarity* dengan 10 cluster .Nilai DBI yang didapatkan adalah sebesar -0.501.

5. KESIMPULAN DAN SARAN

A. KESIMPULAN

Berdasarkan hasil penelitian dan pengujian yang telah dapat ditarik kesimpulan sebagai berikut :

1. Tingkat perhitungan DBI (*Davies Bouldin Index*) menggunakan metode *K-Means* tanpa *cosine similarity*, diperoleh nilai DBI (*Davies Bouldin Index*) terkecil pada *cluster* 10 yaitu sebesar -0.957. Sedangkan nilai DBI terbesar pada *cluster* 2 dengan nilai -0.054. Terdapat anggota di tiap-tiap cluster terdiri dari

- Cluster 1 : 7 Kecamatan
- Cluster 2 : 1 Kecamatan
- Cluster 3 : 1 Kecamatan
- Cluster 4 : 4 Kecamatan
- Cluster 5 : 2 Kecamatan
- Cluster 6 : 1 Kecamatan
- Cluster 7 : 4 Kecamatan
- Cluster 8 : 9 Kecamatan
- Cluster 9 : 10 Kecamatan
- Cluster 10 : 7 Kecamatan
- Total Kecamatan : 46

2. Tingkat perhitungan DBI (*Davies Bouldin Index*) dengan menggunakan metode K-Means *cosine similarity*, diperoleh nilai DBI (*Davies Bouldin Index*) terkecil pada cluster 4 yaitu sebesar -1.483. Sedangkan nilai DBI terbesar pada 10 cluster dengan nilai -1.239. Anggota tiap-tiap cluster terdiri dari:
 - Cluster 1 : 6 Kecamatan
 - Cluster 2 : 24 Kecamatan
 - Cluster 3 : 9 Kecamatan
 - Cluster 4 : 7 Kecamatan
 - Total Kecamatan : 46
3. Tingkat perhitungan nilai DBI (*Davies Bouldin Index*) yang paling baik dengan menggunakan metode *K-means cosine Similarity* dikarenakan nilai yang diperoleh rendah, semakin rendah nilai DBI (*Davies Bouldin Index*) yang diperoleh maka akan semakin baik metode yang digunakan.

B. SARAN

Adapun beberapa saran yang dapat dijadikan dalam pengembangan penelitian ini adalah :

1. Mengembangkan sistem informasi / aplikasi yang dapat lebih membantu pemerintah mengambil keputusan menangani covid. Karena penelitian ini hanya penerapan metode *cosine similarity*.
2. Menggunakan metode selain *Cosine Similarity* untuk meningkatkan kinerja *K-Means*.

6. DAFTAR PUSTAKA

- Achmad Solichin, A, Khairunisa, K. (2020). Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means.
- Alfina, Tahta. Et al, (2020) "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dang Gabungan Keduanya dalam Cluster Data," *Jurnal Teknik ITS*. Vol. 1, ISSN : 2301-9271.
- Aria, Pingit. (7 april 2020). Globalisasi dan Rantai Pasok Dunia yang Terkunci Pandemi Covid-19.
- Amelia Ayu Anggraini 1 Lutfi Ali Muharom, S. Si, M. si2 (2017) PENGELOMPOKAN
- KECAMTAN MENGGUNAKAN METODE K-MEANS CLUSTER.
- Cahyo, M. U, Anggraini, L. B, Maria Andini c , Hesti Retnosari d, M. Anas Nasrulloh e (2021). Penerapan metode K-means clustering data COVID-19 di Provinsi Jakarta.
- Davies, D. L.; Bouldin, D. W. "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2): 224, 1979.
- Nurjanah, M. Hamdani. Dan Astuti, I. Fitri. 2013. Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) untuk Text Mining. *Jurnal Informatika* Volume 8, Nomor 3.
- Noviyanto 2020, Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita COVID-19 Berdasarkan Negara di Benua Asia.
- Paradigma – *Jurnal Informatika dan Komputer*, Vol. 22, No. 2 September 2020
- Rahimi Fitri1, Arifin Noor Asyikin2, *Jurnal POROS TEKNIK*, Volume 7 No.2, Desember 2015 : 54-105, APLIKASI PENILAIAN UJIAN ESSAY OTOMATIS MENGGUNAKAN COSINE SIMILARITY.
- S. Christina, "Kinerja Cosine Similarity dan Semantic Similarity dalam Pengidentifikasian Relevansi Nomor Halaman pada Daftar Indeks Istilah," di Sentika, 2014.
- Windha Mega Pradnya Duhita 2020, Sistem Informasi STMIK AMIKOM Yogyakarta Ring Road Utara Condong Catur Sleman Yogyakarta 55283.
- X. Wu *et al.*, "Top 10 algorithms in data mining," in *Knowledge and Information Systems*, 2008, vol. 14, no. 1, hal. 1–37.