TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
Vol. 4 No. 2 (Desember 2024) – ***Edisi Khusus SEMNASTIK 2024***
https://doi.org/10.46880/tamika.Vol4No2(SEMNASTIK).pp155-159 e-ISSN: 2828-1276

PERBANDINGAN *K-MEANS* DENGAN *HIERARCHICAL CLUSTERING* UNTUK PENGELOMPOKKAN TINGKAT PENGANGGURAN DI SUMATERA UTARA

Enzelina Feronika Situmorang, Yunita Angelina Sihombing, Evalina Damayanti Samosir, Indra M. Sarkis S.✉
Fakultas Ilmu Komputer, Universitas Methodist Indonesia, Medan, Indonesia
Email: poetramora@gmail.com

ABSTRACT

Unemployment is a crucial issue faced by many countries, including Indonesia. This study compares two data clustering methods, K-Means and Hierarchical Clustering, to group districts/cities in North Sumatra based on the Open Unemployment Rate (OUR). The K-Means method is known for its speed and simplicity in partitioning data into clusters by determining centroids as the central points, while Hierarchical Clustering organizes data into a more complex hierarchy without requiring a predefined number of clusters. The OUR dataset used in this study was obtained from various districts/cities in North Sumatra and processed using statistical software to apply both methods. The results indicate that the K-Means method provides superior clustering quality with a Silhouette Score of 56.50%, compared to Hierarchical Clustering, which obtained a score of 43.69%. These findings suggest that the K-Means method is more effective in identifying unemployment patterns in the region. This insight can serve as a reference for policymakers in formulating more targeted strategies to address unemployment in North Sumatera.

Keyword: K-Means, Hierarchical Clustering, Open Unemployment Rate, Data Clustering, North Sumatera.

ABSTRAK

Pengangguran merupakan salah satu permasalahan krusial yang dihadapi oleh banyak negara, termasuk Indonesia. Penelitian ini membandingkan dua metode pengelompokan data, yaitu K-Means dan Hierarchical Clustering, untuk mengelompokkan kabupaten/kota di Sumatera Utara berdasarkan Tingkat Pengangguran Terbuka (TPT). Metode K-Means dikenal karena kecepatan dan kesederhanaannya dalam membagi data ke dalam sejumlah kluster dengan menentukan centroid sebagai titik pusat, sementara Hierarchical Clustering menyusun data dalam bentuk hierarki yang lebih kompleks tanpa memerlukan penentuan jumlah kluster di awal. Dataset TPT yang digunakan diperoleh dari berbagai kabupaten/kota di Sumatera Utara dan diolah menggunakan perangkat lunak statistik untuk menerapkan kedua metode tersebut. Hasil penelitian menunjukkan bahwa metode K-Means memberikan kualitas pengelompokan yang lebih unggul dengan Silhouette Score sebesar 56,50%, dibandingkan dengan Hierarchical Clustering yang memperoleh skor 43,69%. Hasil ini mengindikasikan bahwa metode K-Means lebih efektif dalam mengidentifikasi pola pengangguran di wilayah tersebut. Temuan ini dapat menjadi acuan bagi pembuat kebijakan dalam merumuskan strategi penanganan pengangguran yang lebih tepat sasaran di Sumatera Utara.

Kata Kunci: K-Means, Hierarchical Clustering, Tingkat Pengangguran Terbuka, Pengelompokan Data, Sumatera Utara.

PENDAHULUAN

Pengangguran adalah kondisi di mana seseorang dalam angkatan kerja ingin bekerja, namun belum berhasil mendapatkan pekerjaan. Pengangguran juga mencakup individu yang berkeinginan bekerja, tetapi tidak menemukan pekerjaan yang sesuai dengan bidangnya (Asyfani et al., 2024). Tingkat Pengangguran Terbuka (TPT) adalah indikator yang menunjukkan persentase angkatan kerja yang aktif mencari pekerjaan tetapi belum mendapatkan kesempatan kerja (Purwanda, 2022). TPT mencerminkan kondisi pasar tenaga kerja di suatu wilayah dan menjadi salah satu ukuran penting dalam menganalisis masalah pengangguran. Di Indonesia, termasuk di provinsi Sumatera Utara, TPT menjadi tantangan signifikan yang perlu diatasi untuk meningkatkan kesejahteraan masyarakat dan mengurangi ketimpangan ekonomi.

Penelitian ini membatasi ruang lingkup pada Provinsi Sumatera Utara, dengan fokus pada pengelompokan kabupaten/kota berdasarkan Tingkat Pengangguran Terbuka (TPT) menggunakan dua metode, yaitu K-Means dan Hierarchical Clustering. Faktor lain yang mempengaruhi pengangguran tidak

dianalisis, dan hanya dua metode clustering yang dibandingkan. Selain itu, K-Means memerlukan penentuan jumlah kluster di awal, sementara Hierarchical Clustering lebih kompleks namun fleksibel. Tujuan penelitian ini adalah untuk menemukan metode yang paling efektif dalam mengelompokkan wilayah berdasarkan TPT, sehingga dapat membantu pemerintah daerah membuat kebijakan pengangguran yang lebih tepat sasaran.

METODE PENELITIAN

Penelitian ini menggunakan metode K-Means dan Hierarchical Clustering untuk mengelompokkan kabupaten/kota di Sumatera Utara berdasarkan tingkat pengangguran terbuka (TPT). Kedua metode ini akan diterapkan pada dataset yang sama untuk membandingkan hasil pengelompokan dan menentukan metode yang paling efektif. K-Means merupakan metode partisi yang membagi data ke dalam sejumlah kluster berdasarkan jarak centroid(Oktaviani et al., 2024)(Yanti Liliana et al., 2022), sedangkan Hierarchical Clustering membentuk kluster secara hierarkis, baik dengan pendekatan agglomeratif (penggabungan) atau divisive (pemecahan) (Indra et al., 2023)(Kusumastuti et al., 2022). Analisis hasil dilakukan untuk menilai kinerja masing-masing metode dalam hal kohesi dan pemisahan kluster.

Data yang digunakan diperoleh dari Kaggle, mencakup informasi tentang tingkat pengangguran terbuka di kabupaten/kota di Sumatera Utara (Polak & Cook, 2021). Dataset ini akan diproses dan dianalisis menggunakan perangkat lunak statistik untuk menerapkan kedua metode clustering dan membandingkan hasilnya.

Framework penelitian ini mengikuti langkah-langkah sistematis dalam pengumpulan, pemrosesan, dan analisis data, bertujuan untuk memberikan panduan yang jelas dalam menerapkan dan mengevaluasi kedua metode clustering, serta memastikan keandalan dan relevansi hasil.

Gambar 1. Framework Penelitian

HASIL DAN PEMBAHASAN

K-Means

K-Means Clustering adalah metode partisi yang membagi data menjadi sejumlah kluster berdasarkan kedekatannya dengan centroid kluster. Proses perhitungan K-Means melibatkan langkah-langkah berikut:

a. Inisialisasi Centroid
Pada tahap ini, dipilih centroid awal untuk setiap kluster. Centroid merupakan titik pusat kluster yang dihitung berdasarkan rata-rata nilai dari seluruh data yang termasuk dalam kluster tersebut.

Tabel 1. Inisialisasi Centroid

Kabupaten	2021	2022	2023
Tapanuli Tengah	7.24	7.97	7.81
Tapanuli Utara	1.54	1.07	1.03

b. Hitung Jarak
Pada tahap ini, dilakukan perhitungan jarak antara setiap titik data dengan centroid dari masing-masing kluster. Jarak tersebut digunakan untuk menentukan keanggotaan setiap titik data terhadap kluster yang memiliki centroid terdekat (Primandana et al., 2020). Dalam pendekatan K-Means, jarak Euclidean digunakan sebagai ukuran jarak untuk menilai kedekatan setiap titik data dengan centroid (Wahyudi & Utami, 2022)(Of et al., 2023), dengan tujuan untuk memastikan bahwa setiap kluster terdiri dari titik-titik yang memiliki jarak terkecil dan kesamaan tertinggi terhadap centroid-nya (E. A. Saputra & Nataliani, 2021).

$$d(b_i, a_t) = \sqrt{\sum_{j=1}^{l} (b_{ij} - a_{tj})^2} \quad \text{.....(1)}$$

Jarak dari Tapanuli Tengah ke Centroid 1

$$d = \sqrt{(7.24-7.24)^2 + (7.97-7.97)^2 + (7.81-7.81)^2} = 0$$

Jarak ke Centroid 2 (Tapanuli Utara)

$$d = \sqrt{(7.24-1.54)^2 + (7.97-1.07)^2 + (7.81-1.03)^2} = \sqrt{(5.70)^2 + (6.90)^2 + (6.78)^2} = \sqrt{32.49 + 47.61 + 45.95} = \sqrt{126.05} \approx 11.23$$

Jarak dari Tapanuli Utara ke Centroid 1 (Tapanuli Tengah)

$$d = \sqrt{(1.54-7.24)^2 + (1.07-7.97)^2 + (1.03-7.81)^2} = \sqrt{(-5.70)^2 + (-6.90)^2 + (-6.78)^2} = \sqrt{32.49 + 47.61 + 45.95} = \sqrt{126.05} \approx 11.23$$

Jarak dari Tapanuli Utara ke Centroid 2

$$d = \sqrt{(1.54-1.54)^2+(1.07-1.07)^2+(1.03-1.03)^2} = 0$$

Hierarchical Clustering

Hierarchical Clustering merupakan metode pengelompokan data yang menyusun data dalam hierarki berdasarkan kesamaan antar elemen (Sibarani et al., 2024). Salah satu pendekatan yang digunakan adalah Single Linkage, Single linkage adalah kaedah pengelompokan yang mengelompokkan objek berdasarkan jarak terdekat antara titik dalam kluster (Mutalib et al., 2023) yang mengukur jarak terdekat antara dua kluster. Objek dengan jarak terdekat dikelompokkan dalam kluster yang sama(Suraya & Wijayanto, 2022). Proses perhitungan Hierarchical Clustering dengan metode Single Linkage melibatkan langkah-langkah berikut:

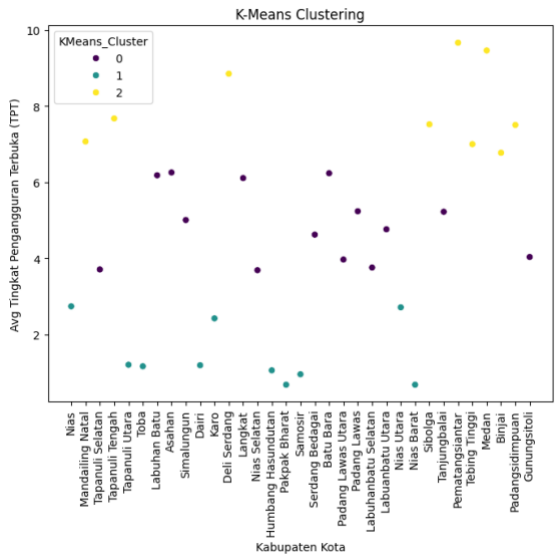
$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\} \quad \text{.....(2)}$$

Hitung jarak Euclidean untuk setiap tahun

$$d_{2021} = \sqrt{(7.24-1.54)^2} = \sqrt{5.7^2} = 5.7$$
$$d_{2022} = \sqrt{(7.97-1.07)^2} = \sqrt{6.9^2} = 6.9$$
$$d_{2023} = \sqrt{(7.81-1.03)^2} = \sqrt{6.78^2} = 6.78$$

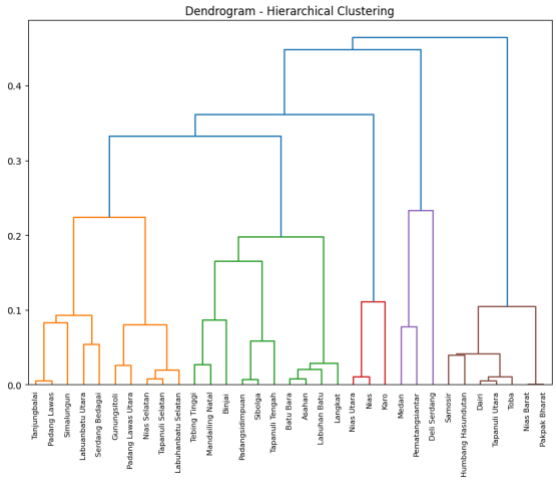
Berikut adalah deskripsi penerapan metode Single Linkage dalam perhitungan jarak antara dua kabupaten:

$$d_{(UV)W} = \min\{d_{2021}, d_{2022}, d_{2023}\}$$
$$d_{(UV)W} = \min\{5.7, 6.9, 6.78\}$$
$$d_{(UV)W} = 5.7$$



Gambar 2. Data Visualization K-Means

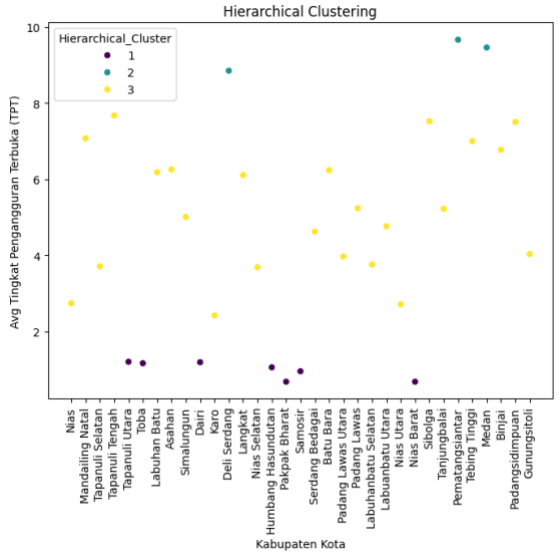
Visualisasi ini menggambarkan hasil klasterisasi K-Means berdasarkan tingkat pengangguran terbuka (TPT) di Sumatera Utara, yang terbagi menjadi tiga klaster. Klaster 0 (Ungu) mencakup kabupaten/kota dengan tingkat pengangguran menengah, seperti Deli Serdang dan Asahan. Klaster 1 (Hijau) berisi kabupaten/kota dengan tingkat pengangguran rendah, contohnya Tapanuli Selatan dan Nias Utara. Sementara itu, Klaster 2 (Kuning) terdiri dari kabupaten/kota dengan tingkat pengangguran tinggi, seperti Kota Medan dan Kota Sibolga. Hasil klasterisasi ini menunjukkan bahwa sebagian besar daerah di Sumatera Utara termasuk dalam Klaster 0, menandakan dominasi tingkat pengangguran menengah di wilayah tersebut.



Gambar 3. Dendogram Hierarchical Clustering

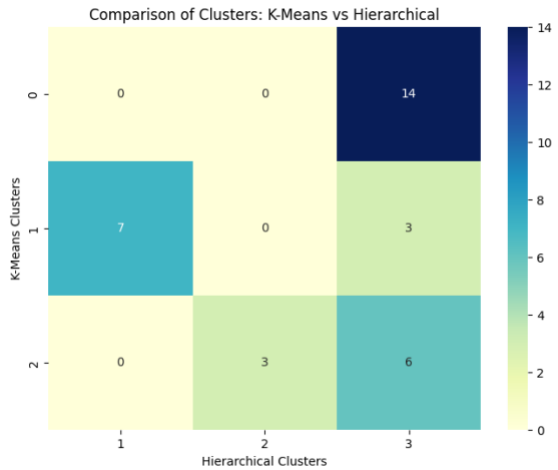
Dendrogram ini menggambarkan hasil klasterisasi hierarkis untuk kabupaten/kota di Sumatera Utara berdasarkan Tingkat Pengangguran Terbuka (TPT). Diagram dua dimensi ini memperlihatkan hubungan antar klaster dan jarak antar klaster tersebut, ditunjukkan oleh panjang garis vertikal yang menghubungkan dua klaster (Asyfani et al., 2024). Dendrogram ini memperlihatkan pengelompokan kabupaten/kota dengan tingkat pengangguran yang serupa, di mana klaster di bagian bawah lebih mirip dan klaster di bagian atas lebih berbeda. Sebagai contoh, Tapanuli Selatan dan Nias Selatan dikelompokkan lebih awal, menunjukkan kesamaan yang tinggi dalam tingkat penganggurannya. Visualisasi ini memberikan gambaran tentang keterkaitan antar wilayah dan dapat mendukung analisis serta perencanaan kebijakan terkait pengangguran

.



Gambar 4. Data Visualization Hierarchical Clustering

Visualisasi ini menampilkan hasil klasterisasi menggunakan metode Hierarchical Clustering berdasarkan rata-rata tingkat pengangguran terbuka (TPT) di kabupaten/kota Sumatera Utara, dengan data terbagi menjadi tiga klaster yang ditandai dengan warna berbeda: ungu (Klaster 1), hijau (Klaster 2), dan kuning (Klaster 3). Klaster 1 (Ungu) mencakup daerah dengan tingkat pengangguran terendah, seperti Nias dan Mandailing Natal, dengan rata-rata TPT antara 1 hingga 2 persen. Klaster 2 (Hijau) berisi satu titik yang mewakili daerah dengan tingkat pengangguran sangat tinggi, mendekati 9 persen. Klaster 3 (Kuning) meliputi sebagian besar daerah dengan TPT bervariasi antara 4 hingga 8 persen, termasuk Kota Medan dan Kabupaten Deli Serdang, menandakan tingkat pengangguran menengah hingga tinggi.



Gambar 5. Komparasi K-Means Vs Hierarchical

K-Means Silhouette Score: 56.50%
Hierarchical Clustering Silhouette Score: 43.69%

Berdasarkan hasil visualisasi heatmap, terdapat perbedaan yang signifikan antara klaster yang dihasilkan oleh metode K-Means dan Hierarchical Clustering. Terlihat bahwa sebagian besar kabupaten/kota di Cluster 3 dari Hierarchical Clustering dikelompokkan ke dalam Cluster 0 oleh K-Means, dengan jumlah tertinggi yaitu 14 kabupaten/kota. Namun, terdapat beberapa kesamaan antara kedua metode, seperti 7 kabupaten/kota yang berada di Cluster 1 pada kedua metode.

Untuk mengevaluasi kualitas pengelompokan, digunakan Silhouette Score. Silhouette Score adalah metrik evaluasi yang mengukur sejauh mana objek dalam suatu klaster berbeda dari klaster lainnya. Nilai Silhouette Score berkisar antara -1 hingga 1, dengan nilai positif menunjukkan bahwa objek lebih cocok dalam klasternya daripada dalam klaster lainnya. Semakin mendekati 1, semakin baik kualitas pengelompokannya (A. Saputra & Yusuf, 2024)(Hendrastuty, 2024).

Hasil analisis menunjukkan bahwa K-Means memiliki kualitas pengelompokan yang lebih baik, dengan Silhouette Score 56,50%, dibandingkan Hierarchical Clustering yang memiliki skor 43,69%. Kedua nilai ini berada dalam rentang yang menunjukkan struktur klaster yang cukup baik (antara 0,25 dan 0,5), namun K-Means menunjukkan kohesi dan separasi yang lebih baik antar klasternya.

Perbedaan Silhouette Score ini juga dapat menjelaskan variasi dalam pengelompokan kabupaten/kota antara kedua metode. K-Means, dengan skor yang lebih tinggi, kemungkinan menghasilkan klaster yang lebih terdefinisi dengan baik, sementara Hierarchical Clustering mungkin menghasilkan batas klaster yang kurang tegas, yang tercermin dalam perbedaan pengelompokan yang diamati.

KESIMPULAN

Penelitian ini membandingkan dua metode pengelompokan, K-Means dan Hierarchical Clustering, dalam mengelompokkan kabupaten/kota di Sumatera Utara berdasarkan tingkat pengangguran terbuka (TPT). Hasil menunjukkan bahwa metode K-Means memiliki kualitas pengelompokan yang lebih baik dengan skor Silhouette sebesar 56,50%, dibandingkan dengan Hierarchical Clustering yang memiliki skor sebesar 43,69%. Meskipun terdapat beberapa kesamaan dalam pengelompokan, K-Means lebih unggul dalam hal kecepatan dan akurasi.

DISEMINASI

Artikel ini telah diseminasikan pada Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASTIK) APTIKOM Tahun 2024 yang

diselenggarakan oleh Universitas Methodist Indonesia pada tanggal 24-26 Oktober 2024.

DAFTAR PUSTAKA

Asyfani, Y., Manfaati Nur, I., Fathoni Amri, I., Yunanita, N., Hikmah Nur Rohim, F., Aura Hisani, Z., & Anggun Lestari, F. (2024). Pengelompokan Kabupaten/Kota di Jawa Tengah Berdasarkan Kepadatan Penduduk Menggunakan Metode Hierarchical Clustering Info Artikel. *Journal of Data Insights*, *2*(1), 1–8.

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (Jima-Ilkom)*, *3*(1), 46–56.

Indra, I., Nur, N., Iqram, Muh., & Inayah, N. (2023). Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting. *INOVTEK Polbeng - Seri Informatika*, *8*(2), 356. https://doi.org/10.35314/isi.v8i2.3612

Kusumastuti, R., Bayunanda, E., Rifa'i, A. M., Asgar, M. R. G., Ilmawati, F. I., & Kusrini, K. (2022). Clustering Titik Panas Menggunakan Algoritma Agglomerative Hierarchical Clustering (AHC). *CogITo Smart Journal*, *8*(2), 501–513. https://doi.org/10.31154/cogito.v8i2.438.501-513

Mutalib, S. S. S. A., Satari, S. Z., & Yusoff, W. N. S. W. (2023). A New Single Linkage Robust Clustering Outlier Detection Procedures for Multivariate Data. *Sains Malaysiana*, *52*(8), 2431–2451. https://doi.org/10.17576/jsm-2023-5208-19

Of, A., & Floyd, T. (2023). Penerapan Algoritma Floyd Warshall dengan Menggunakan Euclidean Distance dalam Menentukan Rute Terbaik. *Jurnal Ilmu Komputer …*, 312–321.

Oktaviani, N., Fauzan, A., & Widyastuti, G. (2024). *Pengelompokan Kabupaten / Kota di Jawa Barat Berdasarkan Tingkat Kesejahteraan Masyarakat Menggunakan K-Means Cluster*. *2*(2), 290–301.

Polak, J., & Cook, D. (2021). A Study on Student Performance, Engagement, and Experience With Kaggle InClass data Challenges. *Journal of Statistics and Data Science Education*, *29*(1), 63–70. https://doi.org/10.1080/10691898.2021.1892554

Primandana, A., Adinugroho, S., & Dewi, C. (2020). Optimasi Penentuan Centroid pada Algoritme K-Means Menggunakan Algoritme Pillar (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi …. *… Teknologi Informasi Dan Ilmu …*, *3*(11), 10678–10683.

Purwanda, E. (2022). The Influence of the Human and Economic Index Development Components on the Unemployment Rate in Indonesia. *Ijd-Demos*, *4*(2), 761–772. https://doi.org/10.37950/ijd.v4i2.264

Saputra, A., & Yusuf, R. (2024). Perbandingan Algoritma DBSCAN dan K-MEANS dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta Menggunakan Metode RFM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1346–1361. https://doi.org/10.57152/malcom.v4i4.1516

Saputra, E. A., & Nataliani, Y. (2021). Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means. *Journal of Information Systems and Informatics*, *3*(3), 424–439. https://doi.org/10.51519/journalisi.v3i3.164

Sibarani, M. A. J. A., Diyasa, I. G. S. M., & Sugiarto, S. (2024). Penggunaan K-Means Dan Hierarchical Clustering Single Linkage Dalam Pengelompokkan Stok Obat. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, *5*(2), 1286-1294

Suraya, G. R., & Wijayanto, A. W. (2022). Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting. *Indonesian Journal of Statistics and Its Applications*, *6*(2), 180–201. https://doi.org/10.29244/ijsa.v6i2p180-201

Wahyudi, A., & Utami, R. (2022). Penggunaan Metode Euclidean Distance Pada Aplikasi Pencarian Lokasi Rumah Sakit di Kota Medan. *Informatics Engineering and Electronic Data (IEED)*, *1*(1), 47–58. https://doi.org/10.59840/ieed.v1i1.193

Yanti Liliana, D., Ermis, I., Zain, A. R., Nurul, D., & Azza, A. (2022). *K-Means Clustering untuk Visualisasi Informasi Pemanfaatan Aplikasi Deteksi Dini Depresi*. *1*(1), 116–123.