Buletin Ilmiah
Ilmu Komputer dan Multimedia

BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

Klastering Penyakit Diabetes dengan Metode K-Means

Muhammad Khalidin Basyir[1*]

[1]Fakultas Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Indonesia
Email: [1*]muhammadkhalidinbasyir@gmail.com
(* : coressponding author)

Abstrak−Penelitian ini bertujuan untuk mengklasifikasikan penyakit diabetes menggunakan metode *K-Means* berdasarkan karakteristik klinis dan demografis pasien. Data yang digunakan diambil dari platform *Kaggle*, terdiri dari 768 entri dengan delapan fitur numerik yang relevan. Proses analisis melibatkan normalisasi data dengan *StandardScaler* dan penentuan jumlah klaster optimal menggunakan metode *Elbow*. Hasil penelitian menunjukkan bahwa tiga klaster optimal dapat diidentifikasi, masing-masing mencerminkan karakteristik unik yang berkaitan dengan profil risiko diabetes. Klaster pertama menunjukkan pasien dengan kadar glukosa tinggi dan BMI tinggi, yang berkaitan dengan diabetes tipe 2. Klaster kedua mencerminkan pasien dengan BMI rendah dan kadar glukosa normal, sedangkan klaster ketiga mengelompokkan pasien dengan kadar insulin rendah dan usia muda. Temuan ini menggarisbawahi potensi algoritma *K-Means* untuk mendukung pengambilan keputusan medis yang lebih personal dan terarah.

Kata Kunci: Diabetes, Klastering, *K-Means*, Analisis Data, Visualisasi

Abstract−*This study aims to classify diabetes using the K-Means Method based on the clinical and demographic characteristics of patients. The data used was taken from the Kaggle platform, consisting of 768 entries with eight relevant numerical features. The analysis process involved normalizing the data with StandardScaler and determining the optimal number of clusters using the Elbow Method. The results showed that three optimal clusters could be identified, each reflecting unique characteristics relating to diabetes risk profiles. The first cluster represents patients with high Glucose levels and high BMI, which is associated with type 2 diabetes. The second cluster reflects patients with low BMI and normal Glucose levels, while the third cluster groups patients with low insulin levels and young Age. These findings underscore the potential of the K-Means algorithm to support more personalized and targeted medical decision-making.*

Keywords*: Diabetes, Clustering, K-Means, Data Analysis, Visualization*

1. PENDAHULUAN

Diabetes adalah salah satu penyakit kronis yang semakin meningkat prevalensinya di seluruh dunia, dengan angka yang mencapai lebih dari 463 juta orang pada tahun 2019 dan diperkirakan akan terus meningkat (Yan et al., 2018). Penyakit ini tidak hanya berdampak pada kesehatan individu, tetapi juga memberikan beban yang signifikan pada sistem kesehatan global. Oleh karena itu, pendekatan yang lebih efektif dalam pengelolaan dan pencegahan diabetes sangat diperlukan. Salah satu metode yang menjanjikan dalam analisis data kesehatan adalah klastering, klustering merupakan metode untuk mengelompokkan titik data menjadi dua kelompok atau lebih, sehingga titik-titik dalam kelompok yang sama memiliki kesamaan yang lebih tinggi dibandingkan dengan titik di kelompok lainnya, dan semua ini dilakukan berdasarkan informasi yang ada pada setiap titik data (Herlinda & Darwis, 2021). Tujuan dari kalstering adalah mengelompokkan karakteristik yang serupa ke dalam satu area, sementara data dengan karakteristik yang berbeda akan tergabung dalam kelompok objek yang memiliki kesamaan (Dewi et al., 2022). Klastering memungkinkan pengelompokan pasien berdasarkan karakteristik klinis dan demografis mereka, sehingga memfasilitasi intervensi yang lebih terarah dan personalisasi dalam perawatan (Yang et al., 2024).

Kajian literatur menunjukkan bahwa berbagai teknik klastering, termasuk *K-Means*, telah diterapkan dalam konteks kesehatan untuk mengidentifikasi pola dalam data pasien dan mendukung pengambilan keputusan klinis (Chen et al., 2017). Sebagai contoh, penelitian oleh Hutchins menunjukkan bahwa teknik klastering dapat digunakan untuk mengidentifikasi tren yang tidak terlihat dalam populasi pasien, yang pada gilirannya dapat membantu dalam merancang intervensi yang lebih sesuai dengan kebutuhan kelompok tertentu. Selain itu, penelitian oleh Yadav menekankan pentingnya teknik klastering dalam pengembangan model prediksi yang dapat meningkatkan layanan kesehatan (Yadav et al., 2023). Namun, meskipun ada kemajuan yang signifikan dalam penerapan teknik ini, masih terdapat kekurangan dalam pemahaman tentang

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

bagaimana klastering dapat dioptimalkan untuk spesifik penyakit diabetes, yang menjadi dasar pernyataan kebaruan ilmiah dari artikel ini.

Pernyataan kebaruan ilmiah dari penelitian ini adalah penerapan metode *K-Means* dalam klastering pasien diabetes dengan mempertimbangkan variabel klinis dan non-klinis yang lebih luas, serta pengembangan model yang dapat memberikan wawasan lebih dalam tentang karakteristik kelompok pasien yang berbeda. Penelitian ini bertujuan untuk mengisi celah dalam literatur yang ada dengan memberikan analisis yang lebih mendalam tentang bagaimana klastering dapat digunakan untuk meningkatkan pemahaman tentang diabetes dan mendukung pengembangan intervensi yang lebih efektif.

Permasalahan penelitian yang diangkat dalam artikel ini adalah bagaimana metode *K-Means* dapat dioptimalkan untuk mengidentifikasi kluster pasien diabetes yang berbeda berdasarkan karakteristik klinis dan demografis mereka. Hipotesis yang diajukan adalah bahwa penerapan metode *K-Means* yang disesuaikan dengan variabel spesifik diabetes akan menghasilkan kluster yang lebih relevan dan informatif, yang pada gilirannya dapat meningkatkan efektivitas intervensi kesehatan yang dirancang untuk kelompok pasien tersebut.

2. METODE PENELITIAN

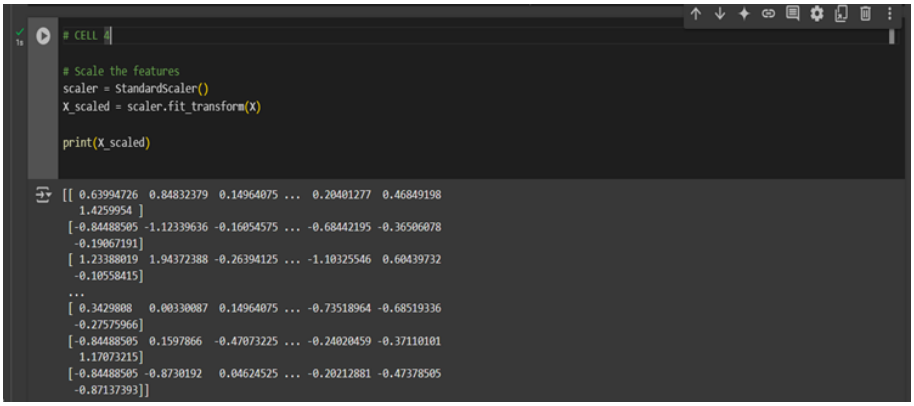
2.1 Persiapan Data

Dataset ini diperoleh dari *Kaggle*, *Kaggle* merupakan sebuah platform yang digunakan oleh berbagai perusahaan untuk menyajikan kompetisi di bidang *data science* (Angkasa & Pangaribuan, 2022). Dataset terdiri dari 768 entri serta 8 fitur numerik, yang digunakan untuk menganalisis penyakit diabetes. Fitur-fitur tersebut mencakup jumlah kehamilan (*Pregnancies*), kadar glukosa dalam darah (*Glucose*), tekanan darah diastolik (*BloodPressure*), ketebalan kulit pada triceps (*SkinThickness*), kadar insulin serum (Insulin), indeks massa tubuh (BMI), fungsi silsilah diabetes (*DiabetesPedigreeFunction*), dan usia pasien (*Age*). Dataset ini tidak mengandung nilai yang hilang, sehingga sangat cocok untuk penelitian pembelajaran mesin, seperti klastering atau prediksi risiko diabetes.

2.2 *Preprocessing*

Tahap *preprocessing* melibatkan normalisasi data menggunakan metode *StandardScaler* dari pustaka *Scikit-learn*. Standard Scaler adalah teknik *preprocessing* yang berfungsi untuk menstandarisasi fitur dengan cara menghilangkan rata-rata dan menskalakan berdasarkan varians (Prasetyo et al., 2022). Normalisasi diperlukan karena fitur dalam dataset memiliki rentang nilai yang berbeda-beda, seperti kadar glukosa yang memiliki nilai jauh lebih besar dibandingkan indeks massa tubuh (BMI). Tanpa normalisasi, fitur dengan skala besar dapat mendominasi hasil klastering.

Setelah normalisasi, setiap fitur memiliki rata-rata 0 dan deviasi standar 1, sehingga memastikan setiap variabel memiliki kontribusi yang seimbang dalam algoritma *K-Means*. Hasil dari proses ini adalah dataset yang siap untuk dianalisis dengan metode klastering.



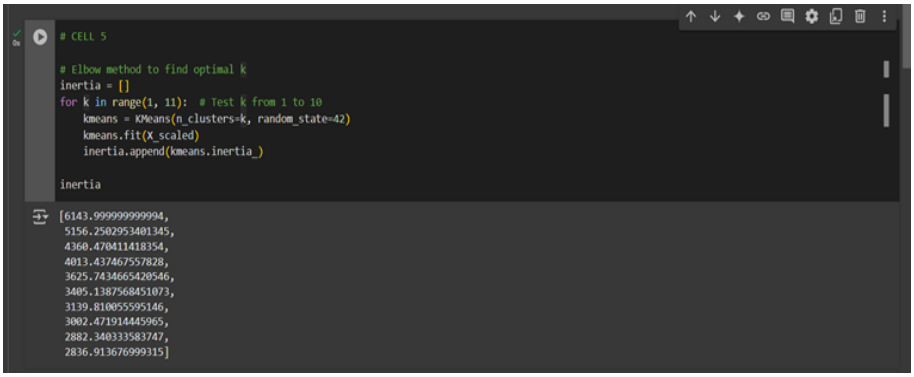
Gambar 1. *Prepocessing* Menggunakan Metode *Standard Scaler*

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

2.3 Penentuan Jumlah *Cluster* (*Elbow Method*)

Metode *Elbow* merupakan salah satu pendekatan yang berfungsi sebagai evaluasi dan digunakan untuk menentukan nilai k optimal dari serangkaian percobaan untuk menguji nilai k (Fuadah et al., 2021). Metode *Elbow* digunakan untuk menentukan jumlah *cluster* (k) yang optimal. Dalam penelitian ini, nilai k diuji dari 1 hingga 10, dan nilai *inertia* dihitung untuk setiap k. *Inertia* adalah total jarak kuadrat antara data dalam *cluster* dan *centorid*nya.

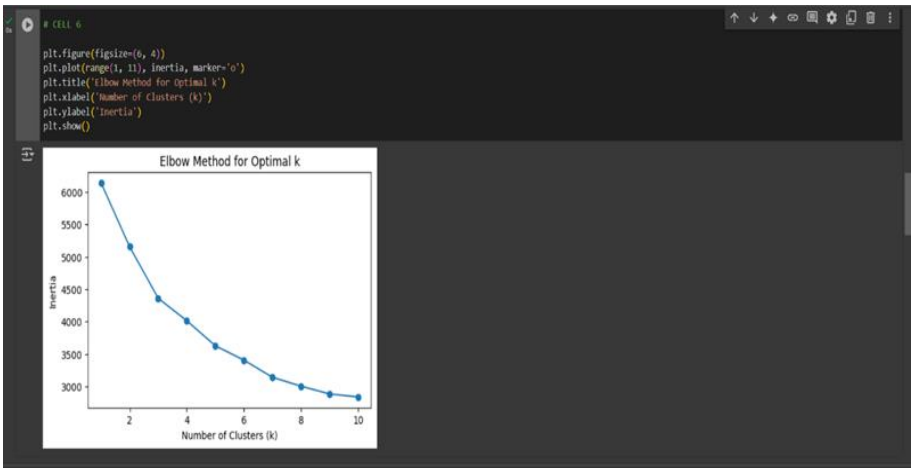
Grafik *Elbow Method* dianalisis untuk menemukan "titik siku" yang menunjukkan penurunan signifikan pada nilai *inertia*. Titik ini mencerminkan jumlah *cluster* optimal, yang merupakan keseimbangan antara variabilitas data yang terjelaskan dan kompleksitas model.



Gambar 2. Penentuan Jumlah Cluster (Elbow Method)

2.3 Penerapan *K-Means* dan Visualisasi

Setelah jumlah *cluster* (k) ditentukan, algoritma *K-Means* diterapkan pada dataset yang telah dinormalisasi. Setiap data kemudian dikelompokkan ke dalam *cluster* berdasarkan jaraknya ke *centorid* terdekat, yang diperbarui secara iteratif hingga konvergen. Proses ini menghasilkan pembagian data menjadi *cluster-cluster* dengan karakteristik yang berbeda.



Gambar 3. Visualisasi *Elbow Method*

3. ANALISA DAN PEMBAHASAN

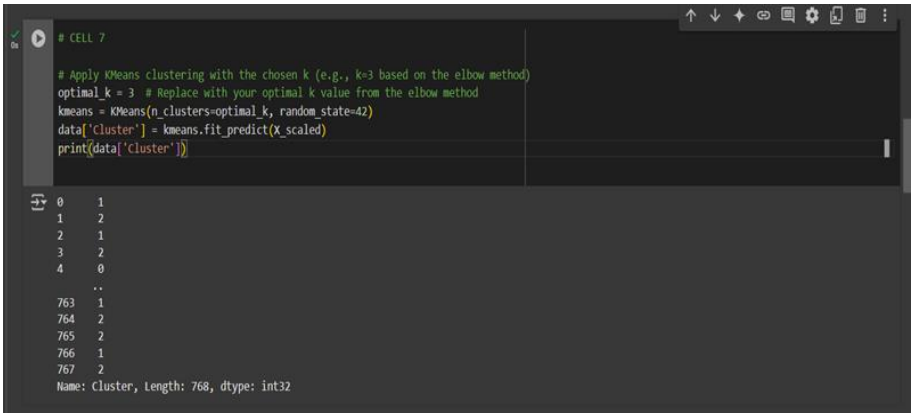
3.1 Penentuan Optimal K

Berdasarkan hasil analisis *Elbow Method*, jumlah *cluster* optimal yang ditemukan adalah k=3. Pada grafik, terlihat penurunan signifikan pada nilai *inertia* antara k=1 hingga k=3, sementara setelah k=3 penurunannya cenderung melandai. Hal ini menunjukkan bahwa k=3 merupakan jumlah

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

cluster yang memberikan keseimbangan terbaik antara variabilitas data yang terjelaskan dan kesederhanaan model.

Pemilihan k=3 juga didukung oleh karakteristik dataset, di mana pembagian ke dalam tiga *cluster* dianggap relevan untuk menggambarkan variasi pasien dengan penyakit diabetes berdasarkan faktor risiko utama seperti kadar glukosa, indeks massa tubuh, dan usia.

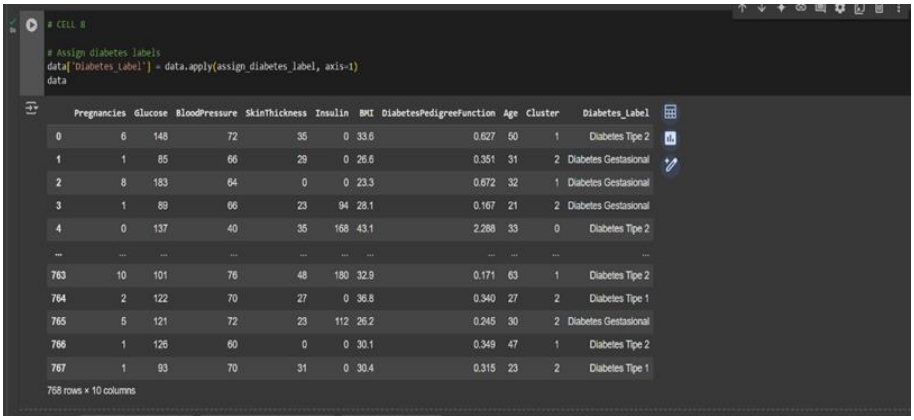


Gambar 4. Penentuan Optimal K

3.2 Klastering Penyakit Diabetes

Hasil klastering menunjukkan bahwa setiap *cluster* memiliki karakteristik yang unik. *Cluster* pertama mencakup pasien dengan kadar glukosa tinggi, BMI tinggi, dan usia di atas 40 tahun, yang sesuai dengan profil pasien diabetes tipe 2. *Cluster* kedua didominasi oleh pasien dengan BMI rendah, kadar glukosa normal, dan usia muda, yang mencerminkan kondisi pasien tanpa risiko diabetes yang signifikan. *Cluster* ketiga mencakup pasien dengan kadar insulin rendah dan usia di bawah 30 tahun, yang lebih cenderung menggambarkan pasien diabetes tipe 1 atau diabetes gestasional.

Distribusi data dalam *cluster* menunjukkan bahwa metode *K-Means* mampu memisahkan kelompok pasien dengan pola karakteristik yang jelas. Temuan ini memberikan wawasan tentang bagaimana faktor risiko tertentu dapat membantu mengidentifikasi tipe diabetes yang berbeda dalam populasi pasien.



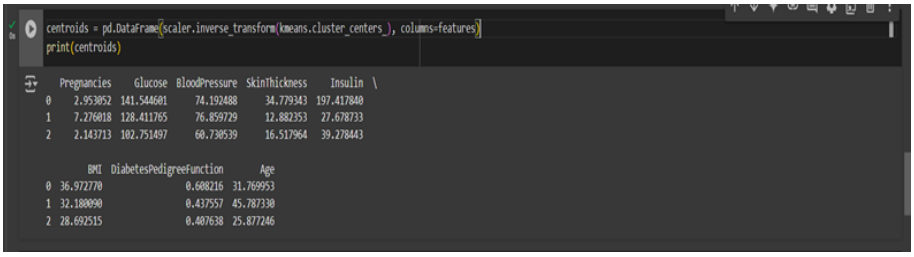
Gambar 5. Klastering Penyakit Diabetes

3.3 Analisis *Centorid* dan Hubungan dengan Tipe Diabetes

Centorid dari setiap *cluster* dianalisis untuk memahami karakteristik utama dari masing-masing kelompok. Sebagai contoh, *centorid cluster* pertama menunjukkan nilai rata-rata glukosa dan BMI yang tinggi, mendukung hipotesis bahwa *cluster* ini didominasi oleh pasien diabetes tipe 2. Sebaliknya, *centorid cluster* kedua memiliki nilai rata-rata glukosa yang rendah, menunjukkan kondisi metabolisme yang lebih sehat.

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

Hasil ini menunjukkan bahwa klastering dapat membantu dalam proses pengelompokan pasien berdasarkan risiko dan karakteristik penyakit diabetes. Jika digabungkan dengan label tipe diabetes, analisis ini dapat memberikan informasi tambahan untuk pengambilan keputusan medis.



Gambar 6. Analisis Centroid

3.4 Visualisasi *Cluster*

Hasil klastering divisualisasikan dalam ruang 3D menggunakan Plotly. Plotly merupakan salah satu pustaka Python yang digunakan untuk menghasilkan grafik dan diagram dalam format tiga dimensi (Rakhmad, 2021). Visualisasi ini memanfaatkan fitur seperti *Glucose*, BMI, dan *Age* untuk memudahkan interpretasi hasil. *Centorid* dari setiap *cluster* juga divisualisasikan untuk memberikan gambaran tentang karakteristik pusat masing-masing *cluster*.

Centorid ditampilkan sebagai tanda silang hitam untuk menandai pusat dari masing-masing *cluster*. Distribusi ini menunjukkan bagaimana algoritma *K-Means* berhasil mengelompokkan data berdasarkan kesamaan fitur.

Melalui visualisasi, dapat diamati bahwa *cluster* dengan nilai glukosa dan BMI yang tinggi cenderung terpisah jauh dari *cluster* dengan nilai glukosa rendah. Hal ini mencerminkan pengaruh signifikan dari faktor-faktor tersebut terhadap pengelompokan pasien diabetes.

Gambar 7. Visualisasi

4. KESIMPULAN

Penelitian ini berhasil menerapkan metode *K-Means* untuk mengelompokkan pasien diabetes berdasarkan karakteristik klinis dan demografis. Berdasarkan analisis *Elbow Method*, jumlah klaster optimal yang ditemukan adalah tiga, dengan setiap klaster menunjukkan pola karakteristik pasien yang berbeda. Klaster pertama mencakup pasien dengan kadar glukosa tinggi, BMI tinggi, dan usia di atas 40 tahun, yang sesuai dengan profil diabetes tipe 2. Klaster kedua mendominasi pasien dengan BMI rendah, kadar glukosa normal, dan usia muda, sedangkan klaster ketiga mencerminkan pasien dengan kadar insulin rendah dan usia di bawah 30 tahun.

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 2, No. 5, Januari 2025
ISSN 3024-8248 (media online)
Hal 904-909

Hasil penelitian ini menunjukkan bahwa metode *K-Means* dapat digunakan secara efektif untuk mengidentifikasi pola dalam data pasien diabetes. Dengan pemahaman yang lebih baik terhadap kelompok risiko, metode ini memiliki potensi besar dalam mendukung pengembangan intervensi kesehatan yang lebih personal dan tepat sasaran. Visualisasi klaster juga memberikan wawasan tambahan yang memperkuat relevansi teknik ini dalam konteks analisis data kesehatan. Penelitian lebih lanjut diperlukan untuk mengintegrasikan klastering dengan label tipe diabetes guna meningkatkan akurasi dan manfaat aplikasinya dalam pengambilan keputusan medis.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah berkontribusi dalam penelitian ini. Terima kasih kepada Universitas Islam Negeri Sumatera Utara, khususnya Program Studi Ilmu Komputer, yang telah memberikan dukungan selama proses penelitian ini berlangsung. Penulis juga mengucapkan terima kasih kepada platform Kaggle yang menyediakan dataset yang digunakan dalam penelitian ini. Selain itu, penghargaan juga disampaikan kepada rekan-rekan sejawat dan pembimbing yang telah memberikan masukan berharga selama proses penyusunan artikel ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan analisis data kesehatan, khususnya dalam pengelolaan penyakit diabetes.

REFERENCES

Angkasa, V., & Pangaribuan, J. J. (2022). Information System Development Komparasi Tingkat Akurasi Random Forest Dan KNN Untuk Mendiagnosis Penyakit Kanker Payudara. *Journal of Information System Development*, *7*(1), 37–38. Retrieved from http://dx.doi.org/10.19166/xxxx

Chen, W., Chen, S., Zhang, H., & Wu, T. (2017). A hybrid prediction model for type 2 diabetes using K-Means and decision tree. *In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 386–390). https://doi.org/10.1109/ICSESS.2017.8342938

Dewi, N. L. P. P., Purnama, I. N., & Utami, N. W. (2022). Penerapan Data Mining Untuk Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: STMIK Primakara). *Jurnal Ilmiah Teknologi Informasi Asia*, *16*(2), 105. https://doi.org/10.32815/jitika.v16i2.761

Fuadah, A. W., Arifin, F. N., & Juwita, O. (2021). Optimasi K-Klasterisasi Ketahanan Pangan Kabupaten Jember Menggunakan Metode Elbow. *INFORMAL Informatics Journal*, *6*(3), 136. https://doi.org/10.19184/isj.v6i3.28363

Herlinda, V., & Darwis, D. (2021). *Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means*. Darwis, Dartono, 2(2), 94–99. Retrieved from http://jim.teknokrat.ac.id/index.php/JTSI

Prasetyo, V. R., Mercifia, M., Averina, A., Sunyoto, L., & Budiarjo, B. (2022). Prediksi Rating Film Pada Website IMDb Menggunakan Metode Neural Network. *Network Engineering Research Operation*, *7*(1), 1. https://doi.org/10.21107/nero.v7i1.268

Rakhmad, K. (2021). *Rekonstruksi Tiga Dimensi Area Pegunungan Berdasarkan Citra Satelit Dua Dimensi Secara Otomatis*. Tugas Akhir - Universitas Islam Indonesia. Retrieved from https://dspace.uii.ac.id/handle/123456789/34042%25

Yadav, S., Singh, M. K., & Kumar, P. (2023). Data Mining Applications for Enhancing Healthcare Services: A Comprehensive Review. *International Journal of Engineering Technology and Management Sciences*, *7*(5), 325–333. https://doi.org/10.46647/ijetms.2023.v07i05.038

Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J., & Low, L. L. (2018). A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology*, *18*(1), 1–12. https://doi.org/10.1186/s12874-018-0584-9

Yang, W. C., Lai, J. P., Liu, Y. H., Lin, Y. L., Hou, H. P., & Pai, P. F. (2024). Using Medical Data and Clustering Techniques for a Smart Healthcare System. *Electronics*, *13*(1), 1–20. https://doi.org/10.3390/electronics13010140