Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

OPTIMASI *K-MEDOIDS* DENGAN PCA UNTUK KLASTERISASI INDIKATOR KESEHATAN IBU HAMIL DI PUSKESMAS LAHOMI

Anisah Hasratniwati Daeli[1], Dian Anubhakti[2*]

[1,2]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia
Email: [1]anisahdaeli@gmail.com, [2*]dian.anubhakti@budiluhur.ac.id
(*: coressponding author)

Abstrak- Tingginya Angka Kematian Ibu (AKI) di Indonesia serta peningkatan kasus Kekurangan Energi Kronik (KEK) ibu hamil di Kabupaten Nias Barat menunjukkan bahwa pemantauan kesehatan ibu hamil masih menghadapi tantangan serius. Di Puskesmas Lahomi, data hasil pemeriksaan kehamilan rutin (*Antenatal Care*/ANC) cenderung hanya dimanfaatkan untuk pelaporan administratif dan belum diolah secara analitik untuk mengidentifikasi pola kesehatan yang lebih mendalam. Permasalahan utama penelitian ini adalah belum optimalnya pemanfaatan data ANC untuk analisis klaster kondisi kesehatan ibu hamil dalam mendukung deteksi dini risiko kesehatan. Penelitian ini bertujuan menganalisis dan membandingkan algoritma *K-Means* dan *K-Medoids* dalam membentuk klaster kondisi kesehatan ibu hamil, serta mengevaluasi pengaruh *Principal Component Analysis* (PCA) sebagai teknik reduksi dimensi dalam meningkatkan kualitas klasterisasi. Data yang digunakan berupa data sekunder hasil pemeriksaan ibu hamil periode Oktober 2023 hingga Maret 2025, dengan variabel utama usia kehamilan, Lingkar Lengan Atas (LILA), dan tekanan darah yang kemudian direkayasa menjadi *Mean Arterial Pressure* (MAP). Pengolahan data dilakukan menggunakan bahasa pemrograman *Python* melalui tahapan pra-pemrosesan, pemodelan, dan evaluasi. Klasterisasi diterapkan pada data dimensi asli dan data hasil reduksi PCA, kemudian dievaluasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index*. Hasil penelitian menunjukkan bahwa *K-Medoids* dengan PCA menghasilkan performa paling optimal pada empat klaster dengan *Silhouette Score* 0,4507 dan DBI 0,7249. Klaster yang terbentuk berhasil mengungkap pola perubahan status gizi serta variasi risiko tekanan darah ibu hamil, yang berpotensi mendukung deteksi dini risiko kesehatan serta pengambilan keputusan berbasis data di Puskesmas Lahomi, serta memberikan kontribusi metodologis dalam penerapan klasterisasi dan reduksi dimensi pada analisis data kesehatan maternal.

Kata Kunci: klasterisasi, k-means, k-medoids, PCA, kesehatan ibu hamil

Abstract- *The high Maternal Mortality Rate (MMR) in Indonesia, along with the increasing cases of Chronic Energy Deficiency (CED) among pregnant women in West Nias Regency, indicates that maternal health monitoring still faces significant challenges. At the Lahomi Health Center, routine Antenatal Care (ANC) data are primarily used for administrative reporting and have not been analytically processed to identify meaningful health patterns. The core problem of this study lies in the limited use of ANC data for clustering maternal health conditions to support early risk detection. This study aims to analyze and compare the performance of K-Means and K-Medoids algorithms in clustering maternal health conditions, as well as to evaluate the effect of Principal Component Analysis (PCA) as a dimensionality reduction technique in improving clustering quality. The dataset consists of secondary data from maternal health examinations conducted between October 2023 and March 2025, with key variables including gestational age, Mid-Upper Arm Circumference (MUAC), and blood pressure, which was further engineered into Mean Arterial Pressure (MAP). Data processing was performed using Python through preprocessing, modeling, and evaluation stages. Clustering was applied to both the original dataset and the PCA-reduced dataset, and performance was assessed using the Silhouette Score and Davies-Bouldin Index. The results show that K-Medoids combined with PCA achieved the most optimal performance with four clusters, obtaining a Silhouette Score of 0.4507 and a DBI of 0.7249. The clusters revealed patterns of nutritional status changes and variations in blood pressure risk among pregnant women, which can support early risk detection and data-driven decision-making for maternal health management at the Lahomi Health Center, while also contributing methodologically to the application of clustering and dimensionality reduction in maternal health data analysis.*

Keywords*: clustering, k-means, k-medoids, PCA, maternal health*

1. PENDAHULUAN

Kesehatan ibu hamil adalah fondasi kualitas kesehatan masyarakat, namun Indonesia masih menghadapi tantangan serius terkait tingginya Angka Kematian Ibu (AKI) yang meningkat menjadi 189 per 100.000 kelahiran hidup pada tahun 2023 [1]. Di tingkat lokal, seperti di Kabupaten Nias Barat, terjadi lonjakan kasus Kekurangan Energi Kronik (KEK) pada ibu hamil dari 219 kasus (2023) menjadi 468 kasus (2024) [2]. Fenomena ini menunjukkan bahwa pemantauan kesehatan ibu di fasilitas layanan primer, seperti Puskesmas Lahomi, belum optimal. Data hasil pemeriksaan kehamilan atau *Antenatal Care* (ANC) yang rutin dikumpulkan cenderung hanya dimanfaatkan untuk pelaporan administratif, bukan untuk analisis mendalam yang dapat mendukung pengambilan keputusan klinis [3].

Untuk mengatasi masalah tersebut, penelitian ini mengusulkan penerapan teknik *data mining*, khususnya *clustering*, untuk mengelompokkan data kesehatan ibu hamil berdasarkan kemiripan indikator klinisnya [4]. Beberapa penelitian terdahulu telah mengaplikasikan metode ini pada kasus kesehatan ibu hamil. Vitara *et al.,* [3]

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

menggunakan metode *K-Means clustering* untuk menganalisis status gizi ibu hamil di Puskesmas Kota Datar, sedangkan Nurtiani dan Astawa [5] menerapkan *K-Means* untuk mengklasifikasikan risiko kesehatan ibu hamil ke dalam empat kelompok risiko. Penelitian lain oleh Riadi *et al.,* [6] mengimplementasikan algoritma C5.0 dan *K-Medoids* untuk menyeleksi atribut paling berpengaruh sebelum mengelompokkan ibu hamil berisiko tinggi, sementara Maulindar dan Yudha [7] mengembangkan pendekatan *clustering* dengan memanfaatkan karakteristik yang relevan untuk mendukung penanganan kesehatan ibu hamil di tingkat layanan dasar. Pendekatan *K-Medoids* dinilai lebih stabil untuk data kesehatan yang mengandung *outlier* karena menggunakan objek aktual sebagai pusat klaster, sebagaimana ditunjukkan oleh Anggraini dan Dzikrullah [8].

Di sisi lain, penelitian Farissa *et al.,* [9] membandingkan *K-Means* dan *K-Medoids* pada data *non-medis* dengan evaluasi *Silhouette Coefficient*. Ishak *et al.,* [10] juga mengoptimalkan *K-Means* pada data penyakit ibu hamil dengan integrasi *Random Forest*, namun penelitian tersebut belum mengintegrasikan teknik reduksi dimensi secara menyeluruh, melainkan sebatas keperluan visualisasi. Padahal, menurut Zebari *et al.,* [11], teknik reduksi seperti *Principal Component Analysis* (PCA) dapat meningkatkan kualitas *clustering* dengan mengeliminasi fitur yang tidak relevan dan menekankan atribut paling berpengaruh.

Meskipun efektif, sebagian besar penelitian klasterisasi pada data kesehatan ibu hamil masih bersifat eksploratif dan deskriptif komparatif, dengan fokus utama pada pemetaan pola atau segmentasi risiko tanpa bertujuan untuk menarik hubungan sebab-akibat antarvariabel. Karakteristik data kesehatan yang kompleks, heterogen, dan sering kali tidak seimbang menjadikan pendekatan klasterisasi relevan sebagai alat bantu pengambilan keputusan berbasis data, namun tidak dimaksudkan untuk menggantikan peran diagnosis klinis [12]. Selain itu, banyak penelitian belum mengevaluasi secara kritis pengaruh pemilihan algoritma dan struktur data terhadap kualitas klaster yang dihasilkan, serta belum mengintegrasikan teknik reduksi dimensi secara sistematis, sehingga potensi redundansi fitur dan *noise* pada data kesehatan ibu hamil masih belum sepenuhnya teratasi.

Dalam konteks data kesehatan ibu hamil, karakteristik data yang bersifat longitudinal berupa pengukuran berulang sepanjang kehamilan menambah kompleksitas proses klasterisasi. Perubahan indikator seperti usia kehamilan, status gizi, dan tekanan darah bersifat dinamis, sehingga diperlukan pendekatan *clustering* yang mampu menangkap pola perkembangan kondisi kesehatan secara konsisten. Penelitian Zhu *et al.,* [13] menunjukkan bahwa *Mean Arterial Pressure* (MAP) memiliki pola perubahan yang relatif konsisten sepanjang gestasi dan berperan penting dalam identifikasi risiko kehamilan. Oleh karena itu, pemilihan algoritma yang stabil serta dukungan teknik reduksi dimensi menjadi aspek penting dalam analisis data kesehatan maternal.

Berdasarkan tinjauan tersebut, terdapat beberapa kesenjangan penelitian. Pertama, sebagian besar penelitian terdahulu hanya mengandalkan satu algoritma *clustering*, terutama *K-Means*, tanpa melakukan perbandingan yang komprehensif dengan metode alternatif seperti *K-Medoids*. Kedua, integrasi teknik PCA sebagai reduksi dimensi dalam klasterisasi data kesehatan ibu hamil masih jarang dilakukan. Ketiga, sebagian besar penelitian menggunakan *dataset* publik atau data umum, sedangkan penelitian ini memanfaatkan data riil hasil pemeriksaan berulang ibu hamil di Puskesmas Lahomi.

Untuk menjawab kesenjangan tersebut, penelitian ini bertujuan menganalisis dan membandingkan performa algoritma *K-Means* dan *K-Medoids* pada data asli dan data hasil reduksi PCA. Analisis ini bertujuan untuk menentukan kombinasi metode dan struktur data yang paling optimal untuk segmentasi kondisi kesehatan ibu hamil di Puskesmas Lahomi. Evaluasi performa model akan diukur menggunakan metrik *Silhouette Score* untuk menilai kepadatan dan pemisahan klaster, serta *Davies-Bouldin Index* (DBI) untuk mengevaluasi rasio sebaran intra-klaster terhadap separasi antar-klaster.

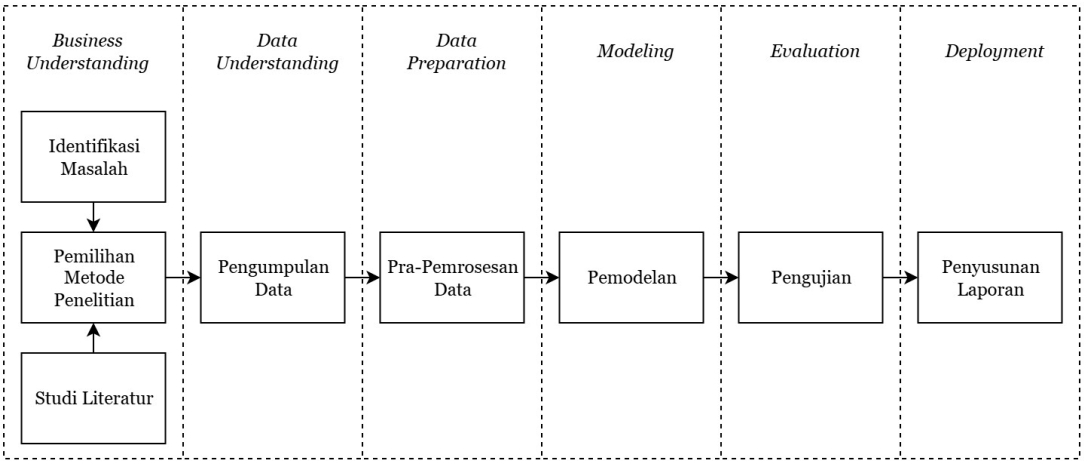
Penelitian ini dibatasi pada penggunaan data sekunder pemeriksaan ibu hamil di Puskesmas Lahomi periode Oktober 2023 hingga Maret 2025, dengan variabel usia kehamilan, Lingkar Lengan Atas (LILA), dan tekanan darah yang direkayasa menjadi *Mean Arterial Pressure* (MAP), serta analisis menggunakan algoritma *clustering K-Means* dan *K-Medoids* yang dievaluasi dengan *Silhouette Score* dan *Davies-Bouldin Index.*

Secara teoretis, penelitian ini memperluas pemahaman mengenai penerapan klasterisasi pada data kesehatan ibu hamil melalui analisis komparatif algoritma *K-Means* dan *K-Medoids* serta pemanfaatan *Principal Component Analysis* (PCA) sebagai teknik reduksi dimensi. Hasil penelitian secara praktis diharapkan dapat menghasilkan klaster yang lebih representatif dalam menggambarkan variasi kondisi kesehatan ibu hamil, sekaligus memberikan kontribusi nyata dalam mendukung deteksi dini risiko kesehatan dan pengambilan keputusan berbasis data di Puskesmas Lahomi.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Cross-Industry Standard Process for Data Mining* (CRISP-DM) sebagai acuan tahapan penelitian. Keseluruhan tahapan penelitian ini dilakukan menggunakan *Jupyter Notebook* dengan bahasa pemrograman *Python*. Proses penelitian dimulai dari identifikasi masalah, studi literatur, pemilihan metode penelitian, pengumpulan data, pra-pemrosesan data, pemodelan, evaluasi, hingga penyusunan laporan. Diagram alir tahapan penelitian disajikan pada Gambar 1.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 1. Diagram Alir Metode Penelitian

2.1 Sumber dan Karakteristik Data

Data penelitian merupakan data sekunder yang bersumber dari rekam medis layanan *Antenatal Care* (ANC) di Puskesmas Lahomi, Kabupaten Nias Barat Periode Oktober 2023 hingga Maret 2025. Data bersifat longitudinal karena memuat pemeriksaan berulang pada individu yang sama sepanjang masa kehamilan. Untuk menjaga kerahasiaan pasien, data telah melalui proses anonimisasi dengan menghapus informasi pribadi dan memberikan kode identifikasi unik untuk setiap individu. Variabel yang dianalisis meliputi usia kehamilan, Lingkar Lengan Atas (LILA), dan tekanan darah yang direkayasa menjadi *Mean Arterial Pressure* (MAP). Variabel-variabel ini dipilih karena merepresentasikan indikator utama status kehamilan, status gizi, dan kondisi kardiovaskular ibu hamil. Deskripsi fitur-fitur yang digunakan pada penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Atribut Analisis Utama

Nama Atribut	Tipe Data	Deskripsi
Usia Kehamilan	Numerik	Usia *gestasi* janin dalam minggu.
TD	Objek	Tekanan darah ibu hamil yang diukur dalam mmHg.
BB	Numerik	Berat badan ibu hamil saat pemeriksaan, diukur dalam kilogram (kg).
LILA	Numerik	Ukuran lingkar lengan atas sebagai indikator status gizi, diukur dalam sentimeter (cm).

2.2 Tahapan Pra-Pemrosesan Data

Pra-pemrosesan data dilakukan untuk mengubah data mentah menjadi *dataset* yang bersih, konsisten, dan siap digunakan dalam analisis [7]. Tahapan ini mencakup penggabungan data, ekstraksi atribut utama, pembersihan data (penanganan karakter spesial, ekstraksi kolom, penanganan nilai kosong dan duplikat, serta penanganan nilai tidak valid dan pencilan), rekayasa fitur, seleksi fitur, dan transformasi data.

2.2.1 Integrasi Data dan Ekstraksi Atribut Utama

Proses integrasi data dilakukan guna menggabung tiga *file* laporan hasil pemeriksaan ibu hamil di Puskesmas Lahomi yang terbagi dalam tiga periode pelaporan yakni Oktober-Desember 2023, Januari- Desember 2024, dan Januari-Maret 2025. Tiga *file Excel* hasil pemeriksaan ibu hamil periode Oktober 2023 hingga Maret 2025 digabungkan menjadi satu *dataframe* utama menggunakan fungsi *concat()* pada *pandas*.

Dari *dataset* yang telah digabungan tersebut, dilakukan ekstraksi empat atribut utama yang relevan untuk analisis , yaitu Usia Kehamilan, Tekanan Darah (TD), Berat Badan (BB), dan Lingkar Lengan Atas (LILA).

2.2.2 Pembersihan Data

Setelah atribut utama diekstraksi, selanjutnya dilakukan tahap pembersihan data, meliputi penyeragaman format, penghapusan data kosong atau duplikasi, serta eliminasi pencilan (*outlier*) berdasarkan validitas klinis untuk meningkatkan kualitas analisis [3]. Inkonsistensi format ditangani dengan menyeragamkan data pada kolom 'Usia Kehamilan' menjadi format numerik dan mengekstrak kolom tekanan darah 'TD' yang berformat teks (contoh: '100/70') menjadi dua kolom numerik terpisah, yaitu 'TD_Sistolik' dan 'TD_Diastolik' menggunakan fungsi *pop()* dari modul *pandas*.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Selanjutnya, integritas data dipastikan dengan menghapus baris yang mengandung nilai kosong dan data yang terduplikasi menggunakan fungsi *dropna()* dan *drop_duplicates()*. Proses ini secara signifikan akan mengurangi jumlah entri data.

Penanganan pencilan (*outlier*) dilakukan dengan pendekatan berbasis validitas klinis. Meskipun deteksi awal menggunakan metode statistik *Z-score* , nilai yang secara klinis valid seperti tekanan darah tinggi (misalnya sistolik 140.0) atau status gizi baik (LILA 30.0) tetap dipertahankan. Hanya nilai yang secara logis tidak mungkin terjadi, seperti berat badan 533 kg atau tekanan diastolik 6.0, yang dihapus dari *dataset*.

2.2.3 Rekayasa Fitur

Selanjutnya, pada tahap rekayasa fitur, atribut tekanan darah sistolik dan diastolik digabungkan menjadi satu variabel baru, yaitu *Mean Arterial Pressure* (MAP). Variabel ini dianggap lebih representatif dalam menggambarkan tekanan darah arteri rata-rata selama satu siklus jantung. Perhitungan MAP menggunakan rumus standar sebagaimana digunakan dalam penelitian longitudinal oleh Zhu *et al.,* [13]:

$$\text{MAP} = \text{DBP} + \frac{1}{3} \times (\text{SBP} - \text{DBP}) \quad (1)$$

Keterangan:
a. MAP : *Mean Arterial Pressure*
b. DBP : *Diastolic Blood Pressure* (Tekanan Darah Diastolik)
c. SBP : *Systolic Blood Pressure* (Tekanan Darah Sistolik)

2.2.4 Seleksi Fitur

Tahap seleksi fitur dilakukan untuk memilih atribut yang paling relevan dengan tujuan analisis. Penelitian terdahulu menunjukkan bahwa pemilihan atribut yang tepat berpengaruh signifikan terhadap kualitas hasil *clustering* [6], [10]. Oleh karena itu, tiga atribut yang paling relevan untuk analisis dipilih: Usia Kehamilan, Lingkar Lengan Atas (LILA), dan *Mean Arterial Pressure* (MAP). Atribut 'Berat Badan' (BB) dieliminasi karena informasinya tumpang tindih dengan LILA dan kurang akurat tanpa data awal kehamilan.

2.2.5 Transformasi Distribusi dan Standarisasi Data

Untuk mengatasi distribusi data yang tidak normal (*skewed*) dan menyeragamkan skala antar fitur, proses transformasi dan standarisasi diterapkan secara simultan menggunakan metode *Box-Cox Transformation* yang terdapat dalam kelas *PowerTransformer*. Metode ini mengubah distribusi data agar mendekati normal sekaligus melakukan standarisasi (*Z-score*), menghasilkan data dengan rata-rata 0 dan standar deviasi 1.

2.2.6 Reduksi Dimensi

Tahap akhir pra-pemrosesan adalah reduksi dimensi dengan metode *Principal Component Analysis* (PCA). Reduksi dimensi bertujuan menyederhanakan data tanpa menghilangkan informasi penting [10]. Teknik ini digunakan untuk mereduksi *dataset* dari tiga dimensi (3D) menjadi dua dimensi (2D). Penerapan PCA dalam klasterisasi terbukti mampu meningkatkan kualitas dan efisiensi pengelompokan data multidimensi, khususnya ketika dikombinasikan dengan algoritma *K-Medoids*, karena menghasilkan klaster yang lebih koheren dan *robust* terhadap *outlier* [14].

2.3 Algoritma Klasterisasi

Pemodelan dilakukan dengan 2 algoritma klasterisasi yakni *K-Means* dan *K-Medoids*.

a. *K-Means adalah* algoritma klasterisasi *non-hirarki* yang mengelompokkan data ke dalam sejumlah klaster (*k*) yang telah ditentukan sebelumnya [5]. Algoritma ini bekerja dengan meminimalkan variasi di dalam klaster dengan cara menghitung jarak setiap objek ke pusat klaster (*centroid*) [15]. Dalam penelitian ini, implementasi *K-Means* dilakukan menggunakan modul *KMeans* dari pustaka *scikit-learn* (*sklearn.cluster*).
b. *K-Medoids*, atau *Partitioning Around Medoids* (PAM), adalah algoritma klasterisasi yang serupa dengan *K-Means* [6]. Perbedaannya, *K-Medoids* menggunakan objek data aktual (*medoid*) sebagai pusat klaster, sehingga lebih tangguh (*robust*) terhadap *noise* dan *outlier* [6], [9]. Pada penelitian ini, algoritma *K-Medoids* diimplementasikan menggunakan *kmedoids* dari pustaka *pyclustering* yang dapat digunakan dengan matriks jarak (*calculate_distance_matrix*).

2.4 Evaluasi Kualitas Klaster

Evaluasi kualitas klaster dilakukan menggunakan dua metrik internal yakni *Silhouette Score* untuk mencari nilai *k* optimal dan *Davies-Bouldin Index* (DBI) sebagai metrik akhir untuk memvalidasi kualitas klaster dari model terbaik yang terbentuk. Penggunaan lebih dari satu metrik evaluasi internal bertujuan untuk memperoleh hasil

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

klasterisasi yang lebih *robust* dan mengurangi bias interpretasi terhadap satu indikator tunggal, sebagaimana direkomendasikan dalam studi komparatif metode klasterisasi [16].

a. *Silhouette Score* digunakan untuk mengevaluasi kualitas klaster dengan mengukur seberapa mirip suatu objek dengan klasternya sendiri dibandingkan dengan klaster lain. Nilai skor berkisar dari -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan kualitas klaster yang sangat baik [17]. Evaluasi ini diimplementasikan menggunakan modul *silhouette_score* dari pustaka *scikit-learn*.
b. *Davies-Bouldin Index* (DBI) digunakan dalam pengujian untuk memvalidasi model terbaik. DBI mengukur rasio kepadatan di dalam klaster dengan pemisahan antar klaster, di mana nilai yang lebih rendah (mendekati nol) menunjukkan kualitas klasterisasi yang lebih baik [10], [17]. Pengujian ini diimplementasikan dengan modul *davies_bouldin_score* dari pustaka *scikit-learn*.

2.5 Desain Eksperimen

Eksperimen dirancang secara sistematis untuk membandingkan pengaruh struktur data terhadap performa algoritma. Disiapkan dua skenario *dataset* untuk tujuan eksperimen dan perbandingan. Pertama adalah *dataset* asli (3D) yang berisi tiga fitur hasil standarisasi (Usia Kehamilan, LILA, dan MAP). Kedua adalah *dataset* hasil reduksi (2D), di mana dataset asli direduksi dimensinya menggunakan *Principal Component Analysis* (PCA) menjadi dua komponen utama (*principal components*). Kedua *dataset* ini digunakan secara paralel dalam tahap pemodelan untuk membandingkan performa algoritma.

2.5.1 Skenario Eksperimen

Menjawab kebutuhan eksperimen yang lebih bervariasi, penelitian ini membagi pengujian ke dalam empat skenario utama seperti terlihat pada Tabel 2.

Tabel 2. Ringkasan Skenario Eksperimen Klasterisasi Data Kesehatan Ibu Hamil

Skenario	Algoritma	Struktur Data	Rentang *k*	Teknik Reduksi Dimensi	Penentuan *k* Optimal
1	K-Means	Data asli	2-5	Tidak diterapkan	*Silhouette Score*
2	K-Medoids	Data asli	2-5	Tidak diterapkan	*Silhouette Score*
3	K-Means	Data hasil reduksi	2-5	PCA	*Silhouette Score*
4	K-Medoids	Data hasil reduksi	2-5	PCA	*Silhouette Score*

Keempat skenario eksperimen dirancang untuk mengevaluasi pengaruh pemilihan algoritma *clustering* dan struktur data terhadap kualitas klaster yang dihasilkan, melalui perbandingan berpasangan antara *K-Means* dan *K-Medoids* pada data asli serta data hasil reduksi dimensi menggunakan PCA. Seluruh eksperimen dilakukan dengan rentang jumlah klaster yang sama dan metrik evaluasi yang konsisten. Jumlah klaster ditetapkan pada rentang k=2 hingga k=5, dengan pertimbangan keseimbangan antara kompleksitas model dan interpretabilitas klaster dalam konteks pelayanan kesehatan primer, serta untuk menghindari *over-segmentation* sebagaimana umum diterapkan dalam penelitian klasterisasi data kesehatan.

2.5.2 Kontrol Eksperimen

Untuk memastikan hasil eksperimen dapat dibandingkan secara adil dan objektif, penelitian ini menerapkan beberapa kontrol eksperimental sebagai berikut:

a. *Dataset*: *Dataset* yang sama digunakan pada seluruh skenario eksperimen dan diterapkan secara konsisten pada kedua algoritma *clustering*.
b. Rentang jumlah klaster (k): Seluruh skenario eksperimen menerapkan rentang jumlah k=2 hingga k=5.
c. Penentuan jumlah klaster optimal: Pemilihan nilai k optimal pada setiap skenario dilakukan menggunakan metrik *Silhouette Score*.
d. *Random State*: Menggunakan nilai *random_state*=42 pada algoritma PCA, *K-Means*, dan *K-Medoids* untuk memastikan hasil yang konsisten pada setiap iterasi.
e. *Environment*: Seluruh pengujian dijalankan pada versi pustaka *scikit-learn* dan *pyclustering* yang sama.

3. HASIL DAN PEMBAHASAN

Penyajian hasil dari proses pengumpulan dan pra-pemrosesan data, serta klasterisasi data kesehatan ibu hamil yang telah dilakukan, disertai pembahasan terhadap pola-pola yang terbentuk dari masing-masing metode yang digunakan, baik dari sudut pandang analitik maupun implikasi metodologisnya.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3.1 Karakteristik *Dataset* Eksperimen

Melalui serangkaian tahapan pengumpulan dan pra-pemrosesan data, *dataset* mengalami perubahan baik dari ukuran, dimensi, maupun representasi nilai. Proses ini menghasilkan dua bentuk representasi data yang digunakan dalam eksperimen klasterisasi. Bentuk pertama adalah *dataset* asli hasil transformasi dan standarisasi data dengan representasi tiga dimensi, yang mencakup variabel Usia Kehamilan, Lingkar Lengan Atas (LILA), dan *Mean Arterial Pressure* (MAP). Bentuk kedua merupakan *dataset* hasil reduksi dimensi menggunakan *Principal Component Analysis* (PCA) yang direpresentasikan dalam dua dimensi.

Dataset asli memiliki 1121 baris data dengan 3 fitur utama, sedangkan *dataset* hasil reduksi PCA mempertahankan jumlah baris yang sama dengan jumlah fitur yang direduksi menjadi 2 komponen utama. Potongan dataset pada kedua representasi ditunjukkan pada Tabel 3 dan Tabel 4.

Tabel 3. Potongan *Dataset* Asli (3 Dimensi)

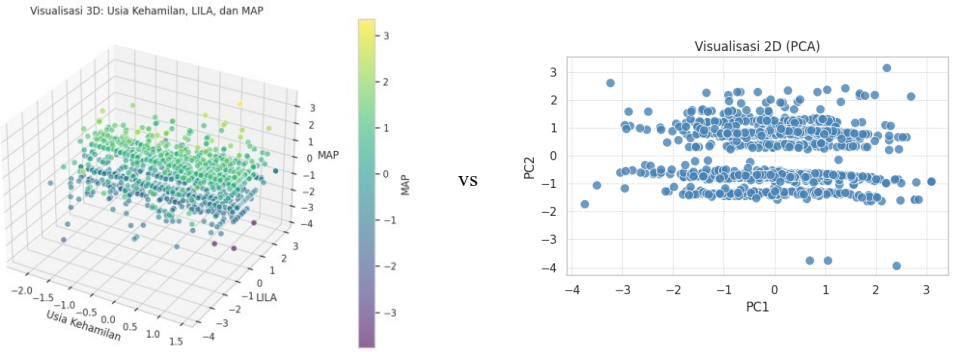
Usia Kehamilan	LILA	MAP
0,6111	-0,4422	0,831
-0,2116	-1,2928	-0,7057
1,4511	-0,4422	0,831
1,1343	-0,4422	0,831
1,4511	0,2891	-1,3386

Data pada Tabel 3 menunjukkan bahwa seluruh fitur telah berada pada skala terstandar, sehingga dapat langsung digunakan sebagai input dalam proses klasterisasi multidimensi.

Tabel 4. Potongan *Dataset* Hasil Reduksi PCA (2 Dimensi)

PC1	PC2
0,1509	0,8943
-1,0874	-0,5892
0,7456	0,9302
0,5213	0,9167
1,1824	-1,2967

Sementara itu, Tabel 4 memperlihatkan hasil transformasi reduksi dimensi data ke dalam dua komponen utama PCA (PC1 dan PC2), yang digunakan sebagai alternatif struktur data dalam eksperimen klasterisasi. Sebaran data dalam ruang tiga dimensi dan dua dimensi divisualisasikan pada Gambar 2.



Gambar 2. Sebaran Data Asli 3D dan Data Reduksi Dimensi PCA 2D

Visualisasi tiga dimensi memberikan gambaran sebaran data berdasarkan ketiga variabel utama dan menunjukkan potensi pembentukan kelompok data. Di sisi lain, visualisasi dua dimensi hasil PCA memperlihatkan distribusi data yang lebih ringkas dan terkompresi, sehingga memudahkan identifikasi struktur global data yang menjadi dasar dalam proses klasterisasi. Secara metodologis, penggunaan dua representasi data ini juga memungkinkan evaluasi pengaruh struktur data terhadap kualitas klaster yang dihasilkan, khususnya dalam konteks data kesehatan dengan korelasi antarvariabel yang cukup kuat.

3.2 Pemodelan

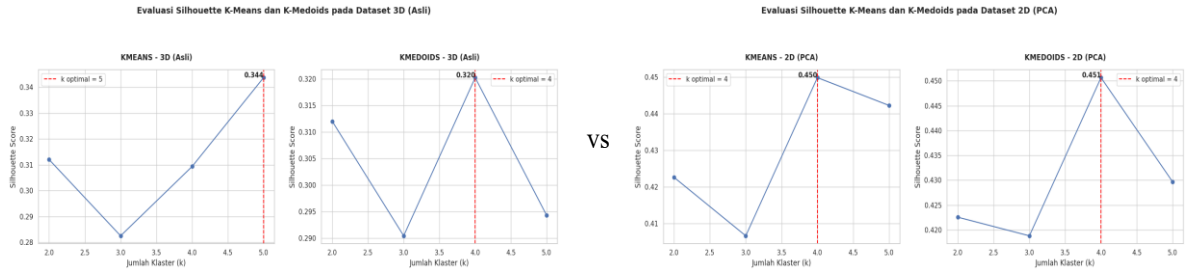
Pemodelan dilakukan untuk membandingkan kinerja algoritma *K-Means* dan *K-Medoids* pada *dataset* asli (3D) dan *dataset* hasil reduksi PCA (2D). Total 16 eksperimen dijalankan dengan memvariasikan jumlah klaster (k) dari 2 hingga 5. Rancangan ini memungkinkan evaluasi sistematis terhadap pengaruh kombinasi algoritma,

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

struktur data, dan jumlah klaster terhadap kualitas pemisahan klaster, sekaligus mengamati konsistensi performa model pada berbagai konfigurasi.

3.2.1 Hasil Komparasi Model

Analisis komparatif dilakukan terhadap performa dua algoritma klasterisasi, yaitu *K-Means* dan *K-Medoids* pada data asli (3D) dan hasil reduksi dimensi dengan PCA (2D). Evaluasi performa model didasarkan pada *Silhouette Score*. yang mengukur sejauh mana objek dalam suatu klaster mirip satu sama lain (kohesi), dan seberapa berbeda dengan objek dalam klaster lain (separasi). Kinerja model pada *dataset* asli (3D) dan *dataset* hasil reduksi dimensi (2D) dapat dilihat pada Gambar 3.



Gambar 3. Evaluasi Kualitas Algoritma pada Kedua Representasi Dataset

Gambar 3 memberikan gambaran awal mengenai perbedaan pola performa algoritma pada masing-masing representasi data, serta menunjukkan kecenderungan peningkatan kualitas klaster setelah penerapan reduksi dimensi PCA. Hasil komparasi dari seluruh eksperimen dan evaluasinya menggunakan *Silhouette Score* dirangkum pada Tabel 5.

Tabel 5. Ringkasan Hasil *Silhouette Score*

Dataset	Metode	k=2	k=3	k=4	k=5
3D (Asli)	K-Means	0.3120	0.2825	0.3093	0.3437
3D (Asli)	K-Medoids	0.3120	0.2884	0.3202	0.3407
2D (PCA)	K-Means	0.4225	0.4066	0.4498	0.4422
2D (PCA)	K-Medoids	0.4225	0.4076	0.4507	0.4297

Hasil pada Tabel 5 menunjukkan bahwa seluruh konfigurasi pada *dataset* hasil reduksi PCA menghasilkan nilai *Silhouette Score* yang lebih tinggi dibandingkan dataset asli, baik pada *K-Means* maupun *K-Medoids*. Hal ini mengindikasikan bahwa reduksi dimensi berkontribusi dalam memperjelas struktur klaster dengan mengurangi redundansi dan korelasi antar fitur. Temuan ini mengindikasikan bahwa reduksi dimensi tidak hanya berfungsi sebagai teknik penyederhanaan visualisasi, tetapi juga berperan penting dalam mengurangi *noise* dan redundansi antar fitur, sehingga struktur klaster menjadi lebih terdefinisi dengan jelas.

Model terbaik secara keseluruhan adalah *K-Medoids* dengan *k*=4 pada *dataset* hasil PCA, yang mencatatkan *Silhouette Score* tertinggi sebesar 0,4507. Secara metodologis, keunggulan *K-Medoids* pada skenario ini menunjukkan ketahanannya terhadap pengaruh *outlier* dan distribusi data yang tidak sepenuhnya homogen, yang merupakan karakteristik umum pada data kesehatan primer.

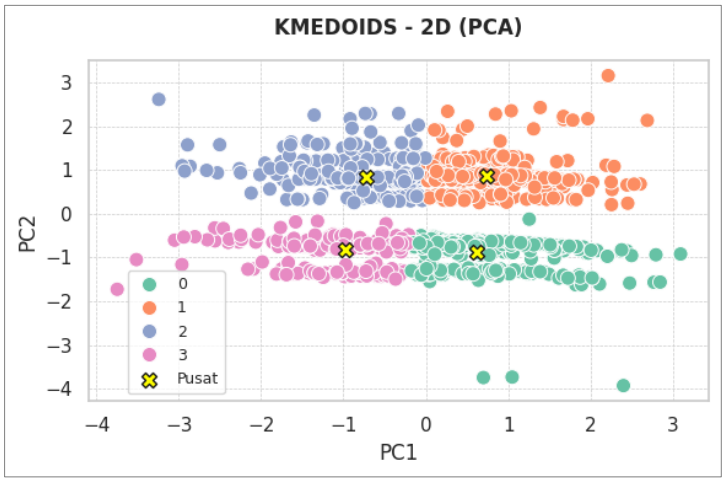
Secara keseluruhan, variasi skenario eksperimen yang melibatkan dua algoritma, dua struktur data, dan beberapa nilai jumlah klaster (*k*) menghasilkan sejumlah temuan metodologis penting. Perbandingan lintas skenario menunjukkan bahwa perubahan struktur data melalui reduksi dimensi PCA memberikan dampak yang lebih signifikan terhadap peningkatan kualitas klaster dibandingkan perbedaan algoritma semata. Selain meningkatkan nilai *Silhouette Score*, PCA juga menghasilkan struktur klaster yang lebih stabil terhadap variasi jumlah klaster. Temuan ini mengindikasikan bahwa pada data kesehatan ibu hamil yang memiliki korelasi antarvariabel cukup kuat, rekayasa struktur data merupakan faktor kunci dalam mengungkap pola klaster yang lebih representatif dan konsisten.

3.2.2 Penyajian dan Analisis Model Terbaik

Setelah dilakukan evaluasi pada seluruh konfigurasi model, *K-Medoids* dengan empat klaster pada data 2D hasil PCA ditetapkan sebagai model terbaik. Model ini tidak hanya memiliki nilai *Silhouette Score* tertinggi, tetapi juga menunjukkan distribusi klaster yang seimbang dan interpretasi yang bermakna secara klinis.

Untuk memastikan bahwa keunggulan model tidak hanya bersifat numerik, tetapi juga mencerminkan struktur klaster yang dapat diinterpretasikan secara visual, dilakukan visualisasi hasil klasterisasi pada ruang dua dimensi PCA.Visualisasi klaster yang terbentuk disajikan pada Gambar 4, di mana setiap titik data diwarnai sesuai kelompoknya dan pusat klaster (*medoid*) ditandai dengan simbol 'X'.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 4. Visualisasi Klaster berdasarkan Model Terbaik

Gambar 4 menunjukkan pemisahan klaster yang relatif jelas dengan tumpang tindih minimal. Hal ini mengindikasikan bahwa proyeksi data ke dalam ruang PCA berhasil mempertahankan informasi penting yang relevan untuk pembentukan klaster. Untuk menginterpretasikan karakteristik masing-masing klaster, dilakukan perhitungan nilai rata-rata dari atribut asli, yakni usia kehamilan, lingkar lengan atas (LILA), dan tekanan darah rata-rata (MAP) pada masing-masing klaster. Hasilnya disajikan pada Tabel 6.

Tabel 6. Karakteristik Klaster Berdasarkan Rata-rata Atribut

Cluster	Jumlah	Proporsi (%)	*Mean* Usia Kehamilan	*Mean* LILA	*Mean* MAP
0	312	27,83	31,88	25,33	68,78
1	295	26,32	33,04	25,2	80,75
2	265	23,64	18,96	24,01	80,98
3	249	22,21	17,37	23,89	69,02

Konfigurasi klaster yang terbentuk menunjukkan bahwa variabel usia kehamilan berperan sebagai sumbu pemisah utama antar klaster, sementara LILA dan MAP berfungsi sebagai variabel pembeda risiko di dalam fase kehamilan yang sama. Hal ini menunjukkan bahwa klasterisasi tidak hanya merefleksikan tahap kehamilan, tetapi juga menangkap heterogenitas kondisi kesehatan ibu hamil pada tahap yang serupa. Hasil analisis dari Tabel 6 menunjukkan pola berikut:

a. Klaster 3 & 2 (Risiko Awal): Klaster ini didominasi oleh ibu hamil pada *trimester* awal hingga pertengahan dengan rata-rata usia kehamilan dan nilai LILA yang lebih rendah dibandinngkan klaster lainnya. Klaster 3 menunjukkan kecenderungan tekanan darah rendah, sedangkan Klaster 2 mencerminkan kelompok dengan MAP yang relatif tinggi. Pola ini mengindikasikan bahwa meskipun status gizi pada fase awal kehamilan masih relatif rendah, variasi tekanan darah mulai muncul dan berpotensi menjadi indikasi awal risiko hipertensi *gestasional*.
b. Klaster 0 & 1 (Fase Lanjutan): Kedua klaster ini merepresentasikan ibu hamil *trimester* akhir dengan rata-rata usia kehamilan dan nilai LILA yang lebih tinggi, yang mencerminkan perbaikan status gizi seiring bertambahnya usia kehamilan. Namun, terdapat perbedaan profil tekanan darah yang signifikan. Klaster 0 menunjukkan nilai MAP yang cenderung lebih rendah, sedangkan Klaster 1 memiliki MAP rata-rata tertinggi, yaitu 80,75 mmHg, yang mengindikasikan kelompok yang memerlukan pemantauan khusus terhadap risiko hipertensi di akhir kehamilan.

Secara keseluruhan, hasil klasterisasi tidak hanya membagi data secara matematis, tetapi juga mengungkap pola dinamis kesehatan ibu hamil sepanjang masa kehamilan. Teridentifikasi adanya tren perbaikan status gizi seiring bertambahnya usia kehamilan, yang kemungkinan dipengaruhi oleh intervensi dan pemantauan rutin. Namun demikian, variasi risiko tekanan darah justru semakin menonjol pada trimester akhir, sehingga menegaskan pentingnya pendekatan pemantauan yang lebih terfokus pada kelompok dengan MAP tinggi.

3.3 Pengujian

Untuk memvalidasi kualitas klaster dari model terbaik, dilakukan pengujian menggunakan metrik *Davies-Bouldin Index* (DBI). Nilai DBI yang lebih rendah menunjukkan bahwa klaster yang terbentuk memiliki kepadatan

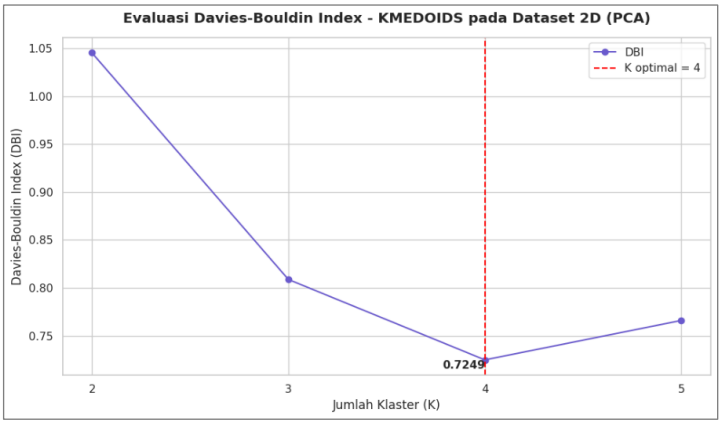
Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

internal yang tinggi dan terpisah dengan baik, sehingga kualitasnya lebih baik. Pengujian dilakukan pada algoritma *K-Medoids* dengan variasi jumlah klaster dari 2 hingga 5. Hasil pengujian disajikan pada Tabel 7.

Tabel 7. Hasil Pengujian dengan Davies-Bouldin Index (DBI)

Jumlah Klaster	*Davies-Bouldin Index* (DBI)
2	1,0452
3	0,8087
4	0,7249
5	0,7660

Interpretasi nilai DBI pada **Error! Reference source not found.** dapat divisualisasikan dalam bentuk grafik garis yang ditunjukkan pada Gambar 5.



Gambar 5. Grafik Nilai DBI pada Variasi Jumlah Klaster

Hasil pengujian pada Tabel 7 dan Gambar 5 menunjukkan bahwa nilai DBI terendah (0,7249) diperoleh saat jumlah klaster adalah 4. Temuan ini konsisten dengan hasil evaluasi *Silhouette Score*, sehingga mengonfirmasi bahwa 4 merupakan jumlah klaster yang paling optimal untuk segmentasi data kesehatan ibu hamil pada penelitian ini.

4. KESIMPULAN

Penelitian ini berhasil menerapkan klasterisasi untuk segmentasi data kesehatan ibu hamil dan menghasilkan beberapa temuan utama. Pertama, kombinasi *K-Medoids* dan PCA terbukti paling optimal dengan *Silhouette Score* tertinggi (0,4507) dan *Davies-Bouldin Index* terendah (0,7249), menunjukkan klaster yang kohesif dan terpisah baik. Kedua, PCA secara signifikan meningkatkan kualitas klasterisasi, terutama pada *K-Medoids* dengan 4 klaster, di mana *Silhouette Score* naik dari 0,3202 menjadi 0,4507. Ketiga, jumlah klaster optimal adalah empat, dengan *K-Medoids* sedikit unggul (0,4507) dibanding *K-Means* (0,4498). Terakhir, klasterisasi berhasil mengungkap pola dinamis kesehatan ibu hamil: status gizi meningkat seiring kehamilan dimana rata-rata LILA naik dari 23,89 cm (*trimester* awal) ke 25,33 cm (*trimester* akhir) dan terdapat variasi risiko tekanan darah di tahap akhir kehamilan, dengan MAP tinggi (80,98 mmHg) dan rendah (69,02 mmHg) pada kelompok berbeda.

Penelitian ini berkontribusi secara metodologis dengan menunjukkan bahwa integrasi PCA dan algoritma *K-Medoids* secara konsisten meningkatkan kualitas klasterisasi data kesehatan ibu hamil dibandingkan pendekatan tanpa reduksi dimensi. Selain itu, hasil penelitian menegaskan bahwa struktur data hasil reduksi dimensi memiliki pengaruh yang lebih signifikan terhadap kualitas klaster dibandingkan pemilihan algoritma semata, sehingga dapat menjadi rujukan metodologis bagi penelitian lanjutan dalam pemodelan dan pemetaan risiko kesehatan maternal berbasis data.

Penelitian ini memiliki beberapa keterbatasan, antara lain penggunaan data yang bersumber dari satu fasilitas layanan kesehatan primer sehingga generalisasi hasil masih terbatas, pembatasan variabel analisis pada tiga indikator klinis utama (usia kehamilan, LILA, dan MAP), serta pendekatan klasterisasi yang bersifat deskriptif dan non-kausal meskipun data yang digunakan bersifat longitudinal.

Berdasarkan keterbatasan tersebut, penelitian lanjutan disarankan untuk melibatkan data dari berbagai wilayah layanan kesehatan, menambahkan variabel klinis dan laboratorium yang lebih beragam, serta mengembangkan pendekatan klasterisasi yang secara eksplisit memodelkan dinamika longitudinal, seperti longitudinal *clustering* atau metode berbasis representasi temporal, guna memperoleh pemetaan risiko kesehatan ibu hamil yang lebih komprehensif dan akurat.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 1, Januari 2026
ISSN 2684-7280 (*online*)
Halaman 103-112
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

DAFTAR PUSTAKA

[1] Kemenkes RI, *Profil Kesehatan Indonesia 2023*. Jakarta: Kementerian Kesehatan RI, 2024.
[2] BPS Kabupaten Nias, *Kabupaten Nias Barat Dalam Angka 2025*, vol. 23. Kabupaten Nias: BPS Kabupaten Nias, 2025.
[3] Hesty Vitara, Rusmin Saragih, and Victor Maruli Pakpahan, "Penerapan Metode Clustering pada Status Gizi Ibu Hamil," *Saturnus J. Teknol. dan Sist. Inf.*, vol. 2, no. 4, pp. 01–16, 2024, doi: 10.61132/saturnus.v2i4.321.
[4] T. Dinh *et al.*, "Data clustering: a fundamental method in data science and management," *Data Sci. Manag.*, 2025, doi: 10.1016/j.dsm.2025.08.001.
[5] N. M. . Nurtiani and I. G. . Astawa, "Penerapan K-Means Clustering Pada Klasifikasi Risiko Kesehatan Ibu Hamil," *J. Nas. Teknol. Inf. dan Apl.*, vol. 1, no. 1, pp. 403–408, 2022, [Online]. Available: https://jurnal.harianregional.com/jnatia/id-92704
[6] M. Riadi, Y. Azhar, and G. W. Wicaksono, "Implementasi Algoritma C5.0 Dan K-Medoids Untuk Klasterisasi Ibu Hamil Beresiko Tinggi," *J. Repos.*, vol. 2, no. 4, pp. 511–524, 2024, doi: 10.22219/repositor.v2i4.30517.
[7] J. Maulindar and E. P. Yudha, "Pengembangan Klastering Untuk Penanganan Ibu Hamil," *Pros. Semin. Nas. Teknol. Inf. dan Bisnis*, pp. 703–708, 2023, [Online]. Available: https://ojs.udb.ac.id/Senatib/article/view/3265
[8] B. S. Kartika Anggraini and Abdullah Ahmad Dzikrullah, "Implementasi Analisis Clustering K-Medoids dalam Pengelompokan Bayi Lahir, Gizi Buruk, dan BBLR Berdasarkan Kecamatan Di Kabupaten Sleman Tahun 2020," *Emerg. Stat. Data Sci. J.*, vol. 2, no. 1, pp. 30–40, 2024, doi: 10.20885/esds.vol2.iss.1.art4.
[9] R. A. Farissa, R. Mayasari, and Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 109–116, 2021, doi: 10.30871/jaic.v5i1.3237.
[10] R. Ishak, N. Nurmawanti, and A. Bengnga, "Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest," *Jambura J. Electr. Electron. Eng.*, vol. 7, no. 1, pp. 41–47, 2025, doi: 10.37905/jjeee.v7i1.28374.
[11] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 1, pp. 56–70, 2020, doi: 10.38094/jastt1224.
[12] A. C. Mawarni, R. Rusdah, L. L. Hin, and D. Anubhakti, "Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest," *IDEALIS Indones. J. Inf. Syst.*, vol. 6, no. 2, pp. 165–171, 2023, doi: 10.36080/idealis.v6i2.3018.
[13] J. Zhu, J. Zhang, N. Syaza Razali, B. Chern, and K. H. Tan, "Mean arterial pressure for predicting preeclampsia in Asian women: a longitudinal cohort study," *BMJ Open*, vol. 11, no. 8, p. e046161, 2021, doi: 10.1136/bmjopen-2020-046161.
[14] U. Hasanah, M. R. Fauziah, A. Fitrianto, E. Erfiani, and L. M. R. D. Jumansyah, "Perbandingan Algoritma Klasterisasi dengan Principal Component Analysis pada Indikator Sosial Ekonomi Kesehatan Jawa Timur," *Techno.Com*, vol. 23, no. 4, pp. 847–863, 2024, doi: 10.62411/tc.v23i4.11534.
[15] E. Abadi, N. Narmawan, S. Umrana, F. Fatmawati, S. Hadrayanti Ananda, and R. Mayangsari, "Pengukuran Antropometri Sebagai Indikasi Kekurangan Energi Kronik pada Ibu Hamil di Wilayah Kerja Puskesmas Puuwatu," *Jukeshum J. Pengabdi. Masy.*, vol. 3, no. 1, pp. 7–11, 2023, doi: 10.51771/jukeshum.v3i1.402.
[16] B. S. Pratama and G. Purwanto, "Perbandingan K-Means dan K-Medoids dalam Pengelompokkan Tingkat Kejahatan pada Provinsi Jawa Tengah," *IDEALIS Indones. J. Inf. Syst.*, vol. 8, no. 2, pp. 295–303, 2025, doi: 10.36080/idealis.v8i2.3562.
[17] I. T. Utami, F. Suryaningrum, and D. Ispriyanti, "K-Means Cluster Count Optimization With Silhouette Index Validation and Davies Bouldin Index (Case Study: Coverage of Pregnant Women, Childbirth, and Postpartum Health Services in Indonesia in 2020)," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 2, pp. 0707–0716, 2023, doi: 10.30598/barekengvol17iss2pp0707-0716.