Volume 7 Nomor 1 Januari 2024 JJEEE e-ISSN : 2715-0887
p-ISSN : 2654-7813

Jambura Journal of Electrical and Electronics Engineering

Optimasi *K-Means* pada *Clustering* Penyakit Ibu Hamil Menggunakan *Random Forest*

Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest

Rezqiwati Ishak*
Progran Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
rezqi.uig@gmail.com*

Nurmawanti
Program Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
nurmawanti237@gmail.com

Amiruddin
Program Studi Teknik Informatika
Universitas Ichsan Goorntalo
Gorontalo, Indonesia
amier.76@gmail.com

Diterima : November 2024
Disetujui : Desember 2024
Dipublikasi : Januari 2025

***Abstrak*— Kesehatan ibu hamil merupakan aspek penting dalam sistem kesehatan masyarakat, di mana pengelompokan data penyakit dapat membantu dalam identifikasi risiko dan perencanaan perawatan yang lebih baik. Namun, metode *clustering* tradisional seperti *K-Means* sering kali menghadapi tantangan dalam pemisahan yang optimal antar *cluster*, terutama ketika atribut yang digunakan tidak relevan. Penelitian ini bertujuan untuk mengoptimalkan metode *K-Means* dalam *clustering* penyakit pada ibu hamil dengan menerapkan seleksi atribut berbasis *Random Forest*. Dari enam atribut yang tersedia (usia, berat badan, tinggi badan, usia kehamilan, sistole, dan diastole), tiga atribut utama yaitu sistole, diastole, dan usia kehamilan dipilih berdasarkan *Importance Score* dari *Random Forest*. Hasil pengujian menunjukkan bahwa penggunaan tiga atribut ini *meningkatkan Silhouette Score* sebesar 0,21 (dari 0,23 menjadi 0,44), yang mengindikasikan pemisahan *cluster* yang lebih baik, serta menurunkan *Davies-Bouldin Index* sebesar 0,69 (dari 1,50 menjadi 0,81), menunjukkan *cluster* yang lebih kompak dan terpisah dengan baik. Visualisasi *clustering* menggunakan *Principal Component Analysis (PCA)* mendukung hasil ini. Selain itu, perhitungan metode *Elbow* menunjukkan jumlah *cluster* optimal pada k=3, memperkuat kesimpulan bahwa pemilihan atribut dan jumlah *cluster* yang tepat meningkatkan kualitas *clustering*. Secara keseluruhan, penelitian ini membuktikan bahwa seleksi fitur berbasis *Random Forest* mampu mengoptimalkan metode *K-Means* dalam *clustering* penyakit pada ibu hamil, yang diharapkan dapat meningkatkan efektivitas diagnosis dan perencanaan perawatan**.

Kata Kunci*— *Clustering; Attribute Selection; Importance Score; Silhouette Score; Davies-Bouldin Index.

Abstract*— *Pregnant women's health is an important aspect of the public health system, where grouping disease data can help in risk identification and better treatment planning. However, traditional clustering methods such as K-Means often face challenges in optimal separation between clusters, especially when the attributes used are irrelevant. This study aims to optimize the K-Means method in disease clustering in pregnant women by applying Random Forest-based attribute selection. Of the six available attributes (age, weight, height, gestational age, systole, and diastole), the three main attributes namely systole, diastole, and gestational age were selected based on the Importance Score from Random Forest. The test results showed that the use of these three attributes increased the Silhouette Score by 0.21 (from 0.23 to 0.44), indicating better cluster separation, and lowered the Davies-Bouldin Index by 0.69 (from 1.50 to 0.81), indicating a more compact and well-separated cluster. Clustering visualization using Principal Component Analysis (PCA) supports these results. In addition, the calculation of the Elbow method shows the optimal number of clusters at k=3, reinforcing the conclusion that the selection of the right attributes and the number of clusters improves the quality of clustering. Overall, this study proves that the selection of Random Forest-based features is able to optimize the K-Means method in disease clustering in pregnant women, which is expected to improve the effectiveness of diagnosis and treatment planning.

Keywords*— *Clustering; Attribute Selection; Importance Score; Silhouette Score; Davies-Bouldin Index.

I. PENDAHULUAN

Kesehatan ibu hamil merupakan salah satu indikator penting dalam sistem kesehatan masyarakat, yang berpengaruh langsung terhadap kesehatan bayi dan perkembangan masyarakat secara keseluruhan. Menurut Organisasi Kesehatan Dunia (WHO), komplikasi selama kehamilan dapat menyebabkan risiko tinggi bagi ibu dan janin, sehingga diperlukan upaya pencegahan dan penanganan yang tepat. Data kesehatan yang akurat dan analisis yang mendalam menjadi kunci dalam mengidentifikasi risiko dan merencanakan intervensi yang efektif[1].

Analisis data kesehatan melalui metode *clustering* telah menjadi alat yang berharga untuk mengelompokkan pasien berdasarkan karakteristik tertentu[2], untuk mendukung pengambilan keputusan yang lebih baik, metode *clustering* atau pengelompokan data menjadi sangat relevan, terutama dalam konteks data kesehatan yang kompleks dan berskala besar.

Berbagai metode *clustering* telah banyak dikembangkan dan digunakan dalam analisis data. Metode yang paling umum adalah *K-Means*[3], *Hierarchical Clustering*[4], dan *DBSCAN*[5]. *K-Means* menjadi salah satu metode yang paling sering digunakan karena kemampuannya dalam mengelompokkan data ke dalam beberapa kelompok berdasarkan kedekatan jarak antar titik data. Selain itu, metode-metode seperti *Hierarchical Clustering* mampu menghasilkan hirarki dalam *clustering*, sedangkan *DBSCAN* efektif dalam mendeteksi *cluster* dengan bentuk yang *arbitrer* dan memiliki kemampuan untuk menangani *noise*[6] [7].

Metode *K-Means*, sebagai salah satu teknik *clustering* yang paling umum digunakan, menawarkan kemudahan dalam pengelompokan data. Namun, *K-Means* memiliki beberapa kelemahan, termasuk ketergantungan pada inisialisasi pusat *cluster* dan sensitivitas terhadap *outlier*, yang dapat mengakibatkan pemisahan *cluster* yang kurang optimal[8][9][10].

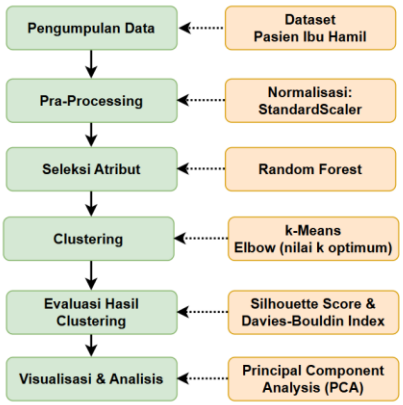
Berdasarkan kelemahan metode *K-Means* di atas, sebagai solusi adalah penerapan seleksi atribut menggunakan algoritma *Random Forest*[11]. Beberapa penelitian sebelumnya yang menggunakan *Random Forest* dalam seleksi atribut atau fitur penting diantaranya oleh[12]. Penelitian ini berfokus pada penerapan *Random Forest* untuk pemilihan fitur dalam konteks data *imunoterapi*. Ini membahas bagaimana pemilihan fitur dapat meningkatkan akurasi klasifikasi sambil menggunakan lebih sedikit fitur, menunjukkan efektivitas *Random Forest* dalam menangani kumpulan data yang tidak seimbang. Selanjutnya penelitian yang dilakukan oleh[13], membahas pentingnya seleksi fitur dalam klasifikasi data, terutama pada dataset dengan banyak variabel. Hasil penelitian ini menunjukkan bahwa *Random Forest* memberikan kinerja terbaik dalam seleksi fitur, menghasilkan akurasi yang lebih tinggi dibandingkan metode lainnya.

Dalam penelitian ini, dilakukan integrasi beberapa metode analisis untuk meningkatkan kualitas *clustering*. Pertama, *StandardScaler* digunakan untuk normalisasi data[14], memastikan bahwa semua atribut memiliki skala yang sama dan mengurangi bias yang mungkin muncul dari perbedaan skala antar atribut. Selanjutnya, *Elbow Method* diterapkan untuk menentukan jumlah *cluster* optimal, memberikan dasar yang lebih kuat untuk pemilihan struktur *cluster* yang tepat[15]. Kualitas *clustering* dievaluasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index*[8], yang memberikan bukti kuantitatif tentang seberapa baik cluster yang dihasilkan. Selain itu, *Principal Component Analysis (PCA)* digunakan untuk reduksi dimensi dan visualisasi hasil *clustering*[16]. Dengan mengurangi jumlah atribut sambil mempertahankan varians data, PCA memungkinkan analisis yang lebih efisien dan memberikan gambaran yang jelas tentang distribusi data dalam ruang yang lebih rendah. Penggunaan PCA tidak hanya memperkuat pemahaman tentang pola kesehatan ibu hamil, tetapi juga membantu dalam mengevaluasi kualitas *clustering* secara visual.

Melalui pendekatan yang komprehensif ini, penelitian ini bertujuan untuk mengoptimalkan metode *K-Means* dalam *clustering* penyakit pada ibu hamil dengan menerapkan seleksi atribut berbasis *Random Forest* dan berbagai teknik evaluasi. Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengelolaan kesehatan ibu hamil, serta meningkatkan efektivitas diagnosis dan perencanaan perawatan.

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan beberapa tahapan, seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Pengumpulan Data
   Tahapan ini dilakukan pengumpulan data dari Dinas Kesehatan Kabupaten Bone Bolango, mencakup atribut usia, berat badan, tinggi badan, usia kehamilan, tekanan darah sistole, tekanan darah diastole, dan riwayat penyakit kehamilan dalam bentuk file Excel.
2. *Pra-Processing*
   Tahapan ini dilakukan normalisasi data dengan *StandardScaler*, data yang telah dikumpulkan akan dinormalisasi menggunakan *StandardScaler* untuk memastikan semua fitur memiliki skala yang sebanding
3. Seleksi Atribut
   Tahapan ini dilakukan seleksi atribut dengan menggunakan *Random Forest*. Sebagian data akan dianalisis menggunakan *Random Forest* untuk menghitung *Feature Importance*. Fitur dengan tingkat kepentingan rendah akan dihilangkan, sementara fitur dengan kepentingan tinggi akan digunakan dalam analisis proses *clustering*.
4. *Clustering* dengan *K-Means*
   Pada tahap setelah seleksi atribut, algoritma *K-Means* akan digunakan untuk mengelompokkan data ke dalam beberapa *cluster*. Jumlah *cluster* yang optimum akan ditentukan menggunakan *Elbow Method*, yang memetakan jumlah *cluster* terhadap *Sum of Squared Errors (SSE)* untuk menemukan titik *"elbow"*.
5. Evaluasi Hasil *Clustering*
   Pada tahap hasil *clustering* akan dievaluasi menggunakan dua metrik utama: *Silhouette Score* dan *Davies-Bouldin Index (DBI)*. *Silhouette Score* mengukur seberapa baik titik data cocok dengan *cluster* yang dihasilkan, sedangkan DBI mengukur rasio antara jarak antar cluster dengan ukuran *cluster*.
6. Visualisasi dengan PCA
   Tahapan ini dilakukan visualisasi hasil *clustering* dengan menggunakan *Principal Component Analysis* (PCA) untuk memproyeksikan data ke dalam dua atau tiga dimensi, memudahkan pemahaman distribusi dan separasi antar *cluster*.

A. Algoritma K-Means

Algoritma *K-Means* adalah salah satu teknik *unsupervised learning* yang paling banyak digunakan untuk *clustering*, di mana tujuan utamanya adalah membagi dataset menjadi beberapa *cluster* berdasarkan kesamaan antar data. Algoritma ini bekerja dengan meminimalkan *within-cluster variance*, yang berarti data dalam satu *cluster* akan memiliki jarak yang lebih dekat satu sama lain dibandingkan dengan data dari *cluster* lain[17][18].

Algoritma *K-Means* menggunakan *Euclidean distance* untuk mengukur jarak antara titik data dengan *centroid*, menggunaan persamaan (1):

$$d(x_i, c_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - c_{jl})^2} \quad (1)$$

di mana:

- $d(x_i,c_j)$ adalah jarak antara titik data x_i dan *centroid* c_j
- x_{il} adalah komponen l dari titik data x_i
- c_{jl} adalah komponen l dari *centroid* c_j,
- n adalah jumlah fitur atau atribut.

B. Standar Scaler

StandardScaler adalah teknik normalisasi yang digunakan untuk memastikan bahwa data memiliki distribusi yang seimbang sebelum diterapkan pada algoritma pembelajaran mesin. Normalisasi sangat penting, terutama ketika data yang digunakan memiliki skala yang berbeda antar fitur. *StandardScaler* adalah metode normalisasi yang berfungsi dengan menstandarisasi fitur agar memiliki *mean nol* dan varian satu[19][20]. *StandardScaler* menggunakan persamaan (2):

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

di mana:

- z adalah nilai data yang telah dinormalisasi,
- x adalah nilai data asli,
- μ adalah rata-rata (*mean*) dari fitur,
- σ adalah standar deviasi dari fitur.

C. Random Forest

Random Forest merupakan metode yang andal untuk mengidentifikasi fitur penting dalam dataset besar dan kompleks. Dengan menggunakan *feature importance*, model dapat difokuskan pada fitur-fitur yang paling relevan, yang dapat mengurangi kompleksitas, meningkatkan akurasi, dan membuat model lebih efisien. Salah satu cara untuk menghitungnya adalah melalui pengurangan *impurity*, yang diukur dengan *Gini Impurity* atau *Entropy*[14][20]. Formula *Gini Impurity* menggunakan persamaan (3):

$$Gini = 1 - \sum_{i=1}^{n} Pi^2 \quad (3)$$

di mana:

- p_i adalah proporsi dari data poin kelas ke- i.
- n adalah jumlah kelas.

Feature importance dihitung berdasarkan seberapa besar pengurangan *impurity* pada setiap node di seluruh pohon dalam hutan. Secara formal, pentingnya fitur f bisa dinyatakan menggunakan persamaan (4):

$$FI(f) = \frac{1}{T} \sum_{t=1}^{T} \Delta I_t(f) \quad (4)$$

di mana:

- FI(f) adalah pentingnya fitur f.
- T adalah jumlah pohon dalam *Random Forest*.
- $\Delta I_t(f)$ adalah pengurangan *impurity* karena fitur f pada pohon ke- t.

D. Metode *Elbow*

Metode *Elbow* adalah teknik yang digunakan untuk menentukan jumlah *cluster* optimal (k) dalam algoritma *K-Means clustering*. Tujuan utama metode ini adalah menemukan nilai k yang menghasilkan pengelompokan terbaik berdasarkan variabilitas dalam data. Metode *Elbow* membantu mengatasi salah satu kelemahan utama dari *K-Means*, yaitu kebutuhan untuk menentukan jumlah *cluster* k terlebih dahulu sebelum algoritma dijalankan[15][21]. *WCSS* (*Within-Cluster Sum of Squares*) menggunakan persamaan (5):

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \epsilon C_i} (x_j - \mu_i)^2 \quad (5)$$

di mana:

- C_i adalah *cluster* ke-i,
- X_j adalah titik data dalam *cluster* C_i,
- μ_i adalah *centroid* dari *cluster* C_i,
- k adalah jumlah *cluster*.

E. Silhouette Score

Silhouette Score adalah metode untuk mengukur kualitas hasil *clustering*, memberikan penilaian seberapa baik data terkelompok. Ini digunakan untuk menentukan seberapa dekat suatu titik data dengan *cluster* yang menjadi anggotanya dibandingkan dengan *cluster* lain. Skor ini berkisar antara -1 hingga 1[17] dengan menggunakan persamaan (6):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

di mana:

- a(i) adalah rata-rata jarak antara titik i dan semua titik lain dalam *cluster* yang sama,
- b(i) adalah rata-rata jarak antara titik i dan titik-titik dalam *cluster* terdekat lainnya (tetangga terdekat),
- s(i) adalah skor *silhouette* untuk titik i.

F. Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) adalah metode evaluasi untuk menilai kualitas hasil *clustering* dengan membandingkan jarak antar *cluster* dengan ukuran dalam-*cluster*, di mana nilai lebih rendah menunjukkan *clustering* yang lebih baik. Nilai DBI memperhitungkan baik jarak antar *cluster* maupun ukuran *cluster*[20] menggunakan persamaan (7):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \quad (7)$$

di mana:

- k adalah jumlah *cluster*,
- S_i adalah ukuran *scatter* dalam *cluster* i, biasanya dihitung sebagai rata-rata jarak antara titik-titik dalam *cluster* i ke *centroid*-nya,
- M_{ij} adalah jarak antara *centroid cluster* i dan j,
- Nilai DB yang lebih rendah menunjukkan *clustering* yang lebih baik.

G. *Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) adalah metode statistik yang digunakan untuk mereduksi dimensi data multivariat tanpa kehilangan informasi yang signifikan. PCA bertujuan untuk menemukan komponen utama *(principal components)* dari sebuah dataset, yang merupakan kombinasi *linier* dari variabel asli dengan tujuan memaksimalkan variasi dalam data. Dalam konteks *clustering*, PCA sering digunakan untuk memproyeksikan data berdimensi tinggi ke dalam dua atau tiga dimensi agar dapat divisualisasikan, sehingga hasil *clustering* menjadi lebih mudah diinterpretasikan[16].

III. HASIL DAN PEMBAHASAN

A. *Pengumpulan Dataset*

Pada tahap ini data yang digunakan berasal dari dataset pasien ibu hamil pada Dinas Kesehatan Kabupaten Bone Bolango yang mencakup berbagai atribut demografis dan kesehatan. Dataset tersebut terdiri dari 300 data yang meliputi informasi penting seperti usia ibu hamil, berat badan, tinggi badan, usia kehamilan, *systol*e (tekanan darah saat denyut jantung), diastole (tekanan darah saat istirahat), dan penyakit yang diderita[22]. Sampel dataset yang digunakan seperti pada Gambar 2.

	No	Usia Ibu Hamil	Berat Badan	Tinggi Badan	Usia Kehamilan	Sistole	Diastole	Penyakit
0	1	28	70	166	37	140	90	0
1	2	35	77	169	38	75	50	1
2	3	23	67	153	36	70	50	1
3	4	23	60	165	9	121	80	2
4	5	24	75	160	36	122	81	3
...	...	...	...	...	...	...	...	...
295	296	37	79	172	19	120	77	23
296	297	30	76	169	18	70	51	1
297	298	19	67	150	32	83	57	1
298	299	25	59	166	19	110	76	3
299	300	30	68	152	20	148	97	4

Gambar 2. Sampel Dataset Pasien Ibu Hamil

B. *Pra-Processing*

Sebelum melakukan analisis lebih lanjut, data yang telah dikumpulkan harus melalui proses *pra-processing* untuk memastikan kualitas dan konsistensi data. Pada tahap ini, dilakukan normalisasi data menggunakan metode *StandardScaler*. Normalisasi ini bertujuan untuk mengubah skala fitur agar setiap atribut memiliki kontribusi yang lebih seimbang dalam analisis. Normalisasi ini menghasilkan nilai yang lebih terpusat di sekitar nol dengan deviasi standar mendekati satu. Hasil normalisasi dengan *StandardScaler* ditampilkan pada Gambar 3.

	Usia Ibu Hamil	Berat Badan	Tinggi Badan	Usia Kehamilan	Sistole	Diastole
0	0.189229	0.009032	0.705086	1.323757	1.245796	0.994956
1	1.168002	0.833699	1.082811	1.441355	-1.687081	-1.423410
2	-0.509894	-0.344397	-0.931721	1.206159	-1.912687	-1.423410
3	-0.509894	-1.169064	0.579178	-1.968976	0.388493	0.390365
4	-0.370069	0.598080	-0.050363	1.206159	0.433615	0.450824
...	...	...	...	...	...	...
295	1.447651	1.069318	1.460536	-0.793000	0.343372	0.208987
296	0.468879	0.715890	1.082811	-0.910597	-1.912687	-1.362951
297	-1.069192	-0.344397	-1.309446	0.735769	-1.326112	-1.000196
298	-0.230245	-1.286873	0.705086	-0.793000	-0.107840	0.148528
299	0.468879	-0.226587	-1.057629	-0.675402	1.606766	1.418170

Gambar 3. Hasil Normalisasi dengan *Standard Scaler*

Proses normalisasi ini penting untuk mengurangi bias yang disebabkan oleh skala atribut yang berbeda, sehingga memudahkan dalam analisis *clustering* dan pengolahan data lebih lanjut.

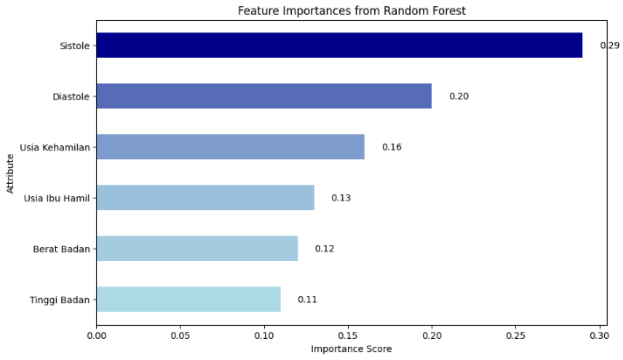
C. *Seleksi Atribut dengan Random Forest*

Setelah tahap *pra-processing* dilanjutkan dengan proses seleksi atribut untuk menentukan atribut-atribut yang paling signifikan dalam pengaruhnya terhadap hasil analisis. Dalam tahap ini, digunakan algoritma *Random Forest* yang memberikan skor pentingnya atribut berdasarkan kontribusi masing-masing fitur terhadap model keseluruhan, sebagaimana hasil perhitungan pada Tabel 1.

TABEL 1. HASIL PERHITUNGAN *RANDOM FOREST*

No	Atribut	*Score Random Forest*
1	Sistole	0,29
2	Diastole	0,20
3	Usia Kehamilan	0,16
4	Usia Ibu Hamil	0,13
5	Berat Badan	0,12
6	Tinggi Badan	0,11

Hasil perhitungan seleksi atribut dengan *Random Forest* pada Tabel 1. Dibuatkan visualisasi seperti pada Gambar 4.

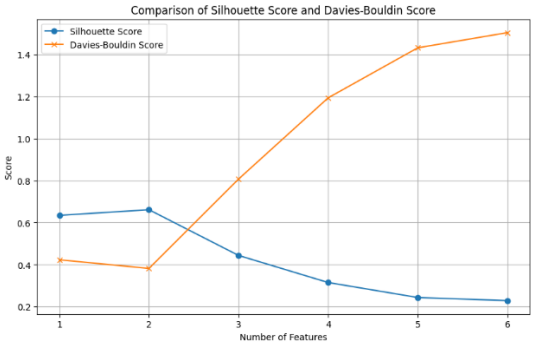


Gambar 4. Hasil Seleksi Fitur dengan *Random Forest*

Hasil analisis menunjukkan bahwa atribut sistole dan diastole memiliki tingkat kepentingan yang tertinggi, dengan skor masing-masing sekitar 0.29 dan 0.20. Selain itu atribut usia kehamilan juga terdeteksi sebagai variabel yang signifikan dan relevan dalam analisis ini, sehingga diputuskan bahwa atribut yang akan digunakan dalam

analisis lebih lanjut adalah sistole, diastole, dan usia kehamilan.

Pemilihan ketiga atribut penting di atas juga didasarkan pada hasil pengujian pengaruh jumlah atribut terhadap kualitas *clustering* dengan mengevaluasi *Silhouette Score* dan *Davies-Bouldin Index*, seperti hasilnya pada Gambar 5.

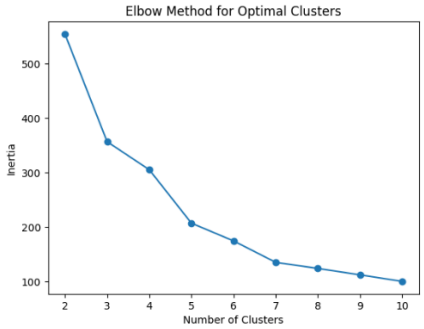


Gambar 5. Pengujian Sejumlah Atribut

Hasil analisis yang ditampilkan pada Gambar 5. menunjukkan bahwa saat jumlah atribut digunakan semakin banyak, *Silhouette Score* justru semakin menurun yang mengindikasikan bahwa pemisahan antar *cluster* menjadi kurang baik. Sebaliknya, nilai *Davies-Bouldin Score* menunjukkan peningkatan seiring bertambahnya jumlah atribut yang menandakan bahwa kualitas pemisahan *cluster* semakin memburuk. Dengan demikian pemilihan atribut yang tepat merupakan langkah kunci untuk mencapai hasil yang optimal dalam *clustering* data pasien ibu hamil, hal ini sejalan dengan penelitian yang dilakukan oleh[11][12][13].

D. Clustrering dengan *K-Means*

Setelah dilakukan seleksi atribut, langkah selanjutnya adalah melakukan *clustering* untuk mengelompokkan data pasien ibu hamil berdasarkan atribut yang telah dipilih. Untuk menentukan jumlah *cluster* (nilai k) yang optimal digunakan metode *Elbow* dengan hasil seperti pada Gambar 6.

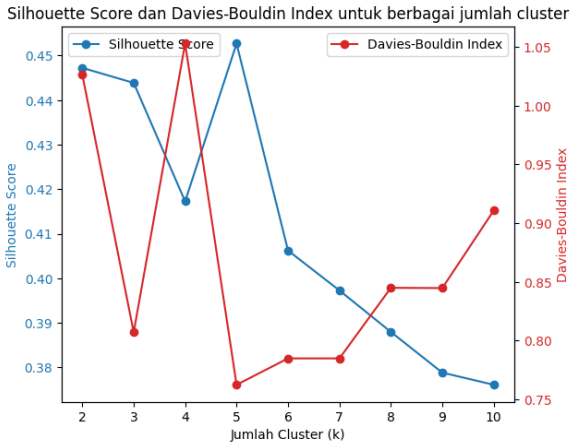


Gambar 6. Penentuan Jumlah *Cluster* Optimal

Meskipun metode *Elbow* menunjukkan bahwa optimal jumlah *cluster* adalah 5, namun peneliti lebih memilih untuk menggunakan k=3 berdasarkan hasil visualisasi yang lebih baik terlihat pada Gambar 10. Visualisasi *clustering* dengan 3 *cluster* memberikan pemisahan yang jelas antar kelompok, memperlihatkan pola yang lebih terstruktur dalam distribusi data pasien. Dengan menggunakan k=3, dapat diidentifikasi karakteristik masing-masing *cluster* yang lebih signifikan.

E. Evaluasi Hasil Clustering

Setelah dilakukan *clustering* dengan k=3, langkah selanjutnya adalah mengevaluasi kualitas hasil *clustering* yang diperoleh dengan menggunakan dua metrik utama untuk evaluasi, yaitu *Silhouette Score* dan *Davies-Bouldin Index*, yang masing-masing memberikan perspektif berbeda tentang sejauh mana *cluster* yang terbentuk terpisah dengan baik, seperti terlihat pada Gambar 7.



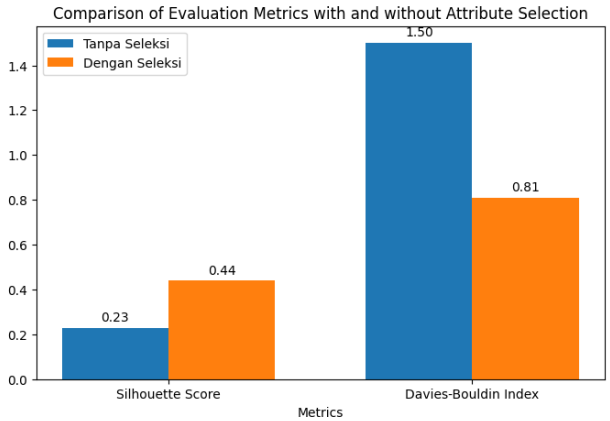
Gambar 7. Evaluasi Hasil *Clustering*

Hasil evaluasi yang ditampilkan pada Gambar 7. menunjukkan nilai *Silhouette Score* dan *Davies-Bouldin Index* untuk berbagai jumlah *cluster* (k). Nilai *Silhouette Score* mencapai puncaknya pada k=3 dan k=5,. Begitu juga untuk *Davies-Bouldin Index*, nilai pada k=3 dan k=5 menunjukkan hasil yang baik, dengan indeks paling kecil, namun pada penelitian ini diputuskan untuk menggunakan jumlah *cluster* k=3 karena setelah dilakukan visualisasi dan analsisi seperti pada Gambar 10. dan Gambar 11. menunjukkan bahwa jumlah *cluster* k=3 hasilnya lebih baik dibandingkan dengan k=5.

Berdasarkan Gambar 5. dapat dianalisis juga bahwa penggunaaan jumlah *cluster* k=3 dengan menggunakan semua atribut (tanpa seleksi atribut) dan hanya menggunakan 3 atribut (hasil seleksi) hasil evaluasinya dapat dilihat pada Tabel 2. dan Gambar 8.

TABEL 2. METRIKS EVALUASI

	Tanpa Seleksi Atribut (6 Atribut)	**Dengan Seleksi Atribut (3 Atribut)**	**Selisih**
Silhouette Score	0.23	0.44	0.21
Davies-Bouldin Index	1.50	0.81	-0.69

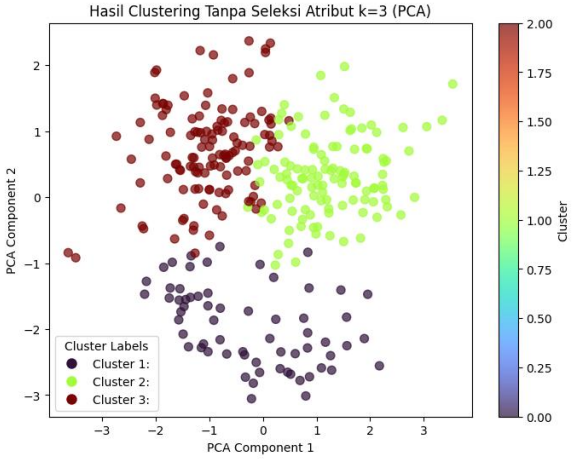


Gambar 8. Metriks Evaluasi

Berdasarkan Tabel 2. dan Gambar 8. menunjukkan bahwa *Silhouette Score* meningkat menjadi 0,44, yang menandakan pemisahan antar *cluster* yang lebih baik, dan *Davies-Bouldin Index* menurun menjadi 0,81, yang menunjukkan *cluster* yang lebih kompak dan terpisah dengan baik sesuai hasil visualisasi hasil *clustering* menggunakan *Principal Component Analysis (PCA)* pada Gambar 9. Dengan demikian penggunaan tiga atribut penting (sistole, diastole, dan usia kehamilan), berhasil meningkatkan optimasi *K-Means* pada *clustering* penyakit ibu hamil. Seleksi fitur dengan *Random Forest* terbukti mampu meningkatkan *Silhouette Score* sebesar 0,21 dan menurunkan *Davies-Bouldin Index* sebesar 0,69, sehingga menghasilkan peningkatan kualitas *clustering* yang signifikan, hal ini sejalan juga dengan penelitian sebelumnya[12][13].

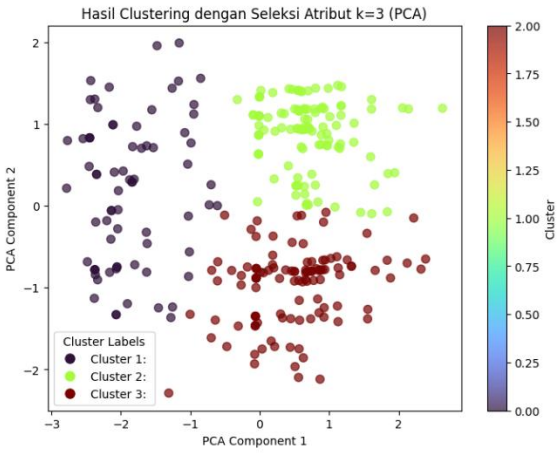
F. Visualisasi dan Analisis

Sebagai langkah akhir dilakukan visualisasi dan analisis hasil *clustering* dengan menggunakan teknik *Principal Component Analysis* (PCA). Hasil visualisasi ini digunakan tiga perbandingan hasil *clustering* yang disajikan pada Gambar 9., Gambar 10. dan Gambar 11.



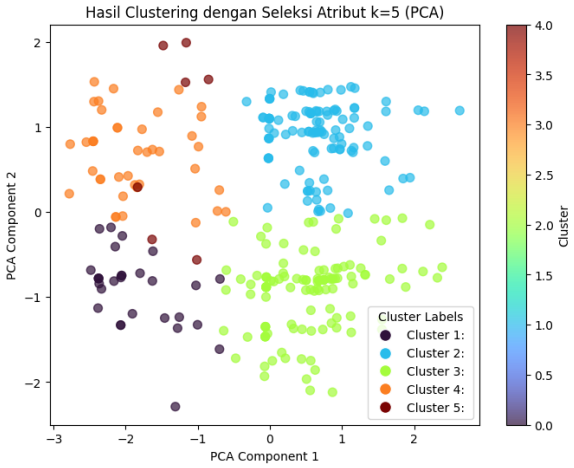
Gambar 9. Hasil *Clustering* Tanpa Seleksi Atribut k=3

Berdasarkan Gambar 9. *clustering* tanpa seleksi atribut dengan k=3 menghasilkan tiga kelompok yang terdefinisi dengan baik, meskipun ada sedikit tumpang tindih antara *Cluster* 1 dan *Cluster* 2.



Gambar 10. Hasil *Clustering* dengan Seleksi Atribut k=3

Beradasarkan Gambar 10. *clustering* dengan seleksi atribut memberikan hasil yang lebih baik, dengan *cluster* yang lebih terisolasi dan terdefinisi. Ini menunjukkan bahwa seleksi atribut membantu dalam mengidentifikasi fitur paling relevan yang berkontribusi pada diferensiasi kelompok setiap *cluster*.



Gambar 11. Hasil *Clustering* dengan Seleksi Atribut k=5

Berdasarkan Gambar 11. Terlihat bahwa dengan lima *cluster*, distribusi data menunjukkan pola yang lebih spesifik di dalam ruang dua dimensi PCA. Setiap *cluster* memiliki posisi yang cukup terpisah, tetapi beberapa *cluster* memiliki posisi yang agak berdekatan, misalnya *cluster* 4 (oranye) dan *cluster* 5 (merah tua), yang mungkin menunjukkan karakteristik yang mirip atau hubungan di antara kedua kelompok tersebut.

Pada *clustering* tanpa seleksi atribut, terdapat lebih banyak tumpang tindih antar *cluster*, yang menunjukkan kurangnya pemisahan yang jelas. Setelah seleksi atribut diterapkan, pemisahan *cluster* menjadi lebih jelas, menunjukkan bahwa seleksi atribut memberikan dampak positif pada diferensiasi antar kelompok.

Cluster menggunakan k=3 dan k=5 dengan seleksi atribut memiliki pemisahan yang lebih baik, meskipun beberapa *cluster* di k=5 terlihat berdekatan. Ini menunjukkan bahwa model *clustering* yang dilengkapi seleksi atribut memberikan hasil yang lebih jelas dan terpisah. Model *clustering* terbaik dalam penelitian ini adalah model dengan seleksi atribut dan jumlah *cluster* k=3.

IV. KESIMPULAN

Pengujian optimasi *clustering* menggunakan *Silhouette Score* dan *Davies-Bouldin Index* menunjukkan peningkatan yang signifikan ketika hanya tiga atribut penting (sistole, diastole, dan usia kehamilan) digunakan, dibandingkan dengan penggunaan enam atribut (usia, berat badan, tinggi badan, usia kehamilan, tekanan sistole, dan diastole). Hasil yang didapatkan adalah (1). Hasil dengan enam atribut, *Silhouette Score* mencapai 0,23, yang menandakan pemisahan antar *cluster* yang kurang optimal. *Davies-Bouldin Index* sebesar 1,50 menunjukkan bahwa *cluster* yang terbentuk kurang kompak., dan (2). hasil dengan tiga atribut, setelah menerapkan tiga atribut penting, *Silhouette Score* meningkat menjadi 0,44, mengindikasikan pemisahan antar *cluster* yang lebih baik. *Davies-Bouldin Index* turun menjadi 0,81, menunjukkan bahwa *cluster* yang terbentuk lebih kompak. Hasil ini juga sejalan dengan visualisasi *clustering* yang

dilakukan menggunakan *Principal Component Analysis (PCA)*. Metode *Elbow* mendukung temuan ini dengan menunjukkan bahwa jumlah *cluster* optimal adalah k=3, yang memperkuat pemilihan tiga atribut ini sebagai faktor kunci dalam optimasi algoritma *K-Means*. Selain itu, seleksi fitur menggunakan *Random Forest* terbukti efektif dalam meningkatkan kualitas *clustering*, dengan *Silhouette Score* meningkat sebesar 0,21 dan *Davies-Bouldin Index* menurun sebesar 0,69. Hal ini menunjukkan bahwa pemilihan atribut yang tepat secara signifikan meningkatkan kualitas hasil clustering.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih yang sebesar-besarnya kepada Ketua Yayasan Pengembangan Ilmu Pengetahuan dan Teknologi (YPIPT) Ichsan melalui Ketua Lembaga Penelitian Universitas Ichsan Gorontalo atas dukungan dana penelitian yang diberikan pada program Penelitian Kompetitif Dosen Unisan (PKDU) tahun 2024. Tanpa bantuan dan kepercayaan yang diberikan, penelitian ini tidak akan dapat terlaksana dengan baik. Kami juga menghargai perhatian dan komitmen yang telah diberikan dalam mendukung pengembangan penelitian di bidang Ilmu Komputer demi kemajuan ilmu pengetahuan dan teknologi. Semoga kerja sama ini dapat terus berlanjut dan bermanfaat bagi kemajuan penelitian di masa yang akan datang.

REFFERENSI

[1] WHO, "Maternal Health," 2021. https://www.who.int/health-topics/maternal-health#tab=tab_1 (accessed Sep. 01, 2024).

[2] Z. M. Kesuma, Nurhasanah, and P. Kesuma, "Maternal health care in Aceh Province: cluster analysis results," *J. Phys. Conf. Ser.*, vol. 1116, p. 022019, Dec. 2018, doi: 10.1088/1742-6596/1116/2/022019.

[3] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, no. 10, pp. 80716–80727, Mar. 2020, doi: 10.1109/ACCESS.2020.2988796.

[4] X. Li, Y. Ye, M. J. Li, and M. Ng, "On cluster tree for nested and multi-density data clustering," *Pattern Recognit.*, vol. 43, pp. 3130–3143, 2010, doi: 10.1016/j.patcog.2010.03.020.

[5] R. V. S. Kumar, "An Efficient Clustering Approach using DBSCAN," *HELIX*, vol. 8, no. 3, pp. 3399–3405, Apr. 2018, doi: 10.29042/2018-3399-3405.

[6] S. K. Majhi and S. Biswal, "Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer," *Karbala Int. J. Mod. Sci.*, vol. 4, no. 4, pp. 347–360, Dec. 2018, doi: 10.1016/j.kijoms.2018.09.001.

[7] N. Gholizadeh, H. Saadatfar, and N. Hanafi, "K-DBSCAN: An improved DBSCAN algorithm for big data," *J. Supercomput.*, vol. 77, no. 6, pp. 6214–6235, Jun. 2021, doi: 10.1007/s11227-020-03524-3.

[8] 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, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0707-0716.

[9] A. Bengnga and R. Ishak, "Penerapan XGBoost untuk Seleksi Atribut pada K-Means dalam Clustering Penerima KIP Kuliah," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 2, pp. 192–196, 2023, doi: 10.37905/jjeee.v5i2.20253.

[10] A. Bengnga and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," *Jambura J. Electr. Electron. Eng.*, vol. 4, no. 2, pp. 169–174, Jul. 2022, doi: 10.37905/jjeee.v4i2.14403.

[11] Z. Wang, H. Li, B. Nie, J. Du, Y. Du, and Y. Chen, "Feature selection using different evaluate strategy and random forests," in *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, Aug. 2021, pp. 310–313, doi: 10.1109/ICCEAI52939.2021.00062.

[12] A. Y. Mahmoud, "Novel efficient feature selection: Classification of medical and immunotherapy treatments utilising Random Forest and Decision Trees," *Intell. Med.*, vol. 10, p. 100151, 2024, doi: 10.1016/j.ibmed.2024.100151.

[13] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[14] S. Rajesh, P. Praveen, and D. N, "Performance Analysis of Machine Learning Algorithms on Parkinson's Disease Data," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Mar. 2024, pp. 1–10, doi: 10.1109/InC460750.2024.10649372.

[15] A. Damayanti, W. D. Utami, D. C. R. Novitasari, P. K. Intan, and M. L. Kurniawan, "Cluster Analysis of Environmental Pollution in Indonesia Using Complete Linkage Method with Elbow Optimization," *JTAM (Jurnal Teor. dan Apl. Mat.*, vol. 7, no. 2, p. 399, Apr. 2023, doi: 10.31764/jtam.v7i2.12961.

[16] I. Turbay, P. Ortiz, and R. Ortiz, "Statistical analysis of principal components (PCA) in the study of the vulnerability of Heritage Churches," *Procedia Struct. Integr.*, vol. 55, pp. 168–176, 2024, doi: 10.1016/j.prostr.2024.02.022.

[17] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2019.

[18] E. Prasetyo, *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: CV. Andi Offset, 2013.

[19] R. Primartha, *Algoritma Machine Learning*. Bandung: Informatika, 2021.

[20] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: :Informatika, 2018.

[21] Y. Heryadi and T. Wahyono, *Machine Learning Konsep dan Implementasi*. Yogyakarta: Gava Media, 2020.

[22] Dinkes, "Laporan Harian Pelayanan Pasien," Gorontalo, 2024.