Identifikasi Pola Risiko Pasien Polycystic Ovary Syndrome (PCOS) Menggunakan Algoritma K-Means

Fany Charoliana Dewi[1*], Melda Agarina[2]
Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya,
Kota Bandar Lampung, Indonesia
Email: fany6954@gmail.com[1], agharina@darmajaya.ac.id[2]

Abstrak

Sindrom Ovarium Polikistik (*Polycystic Ovary Syndrome*/PCOS) merupakan gangguan endokrin kompleks yang banyak dialami wanita usia reproduktif dan berkaitan dengan kondisi reproduksi serta metabolik. Oleh karena itu, diperlukan pendekatan analisis data yang sistematis untuk mengidentifikasi pola risiko PCOS. Penelitian ini bertujuan mengelompokkan data pasien PCOS menggunakan algoritma *K-Means* dan menganalisis pola risiko yang terbentuk. Data penelitian bersumber dari *dataset* publik Kaggle yang terdiri atas 2.539 data pasien dengan variabel demografis, hormonal, metabolik, reproduksi, dan gaya hidup. Metode penelitian mengacu pada tahapan *Knowledge Discovery in Databases* (KDD), yang meliputi *data selection, preprocessing, transformation, data mining,* serta evaluasi dan interpretasi. Algoritma *K-Means* diterapkan untuk mengelompokkan data ke dalam tiga kategori, yaitu risiko rendah, sedang, dan tinggi. Hasil penelitian menunjukkan bahwa pengelompokan menghasilkan kelompok dengan karakteristik yang berbeda berdasarkan kemiripan data. Pola yang terbentuk mencakup karakteristik seperti indeks massa tubuh (BMI), ketidakteraturan siklus menstruasi, dan parameter hormonal. Hasil tersebut memberikan gambaran mengenai pola karakteristik pasien PCOS yang dapat menjadi informasi pendukung bagi tenaga medis dalam memahami data pasien serta meningkatkan pengetahuan masyarakat mengenai pencegahan dan penanganan PCOS.

Kata Kunci: *K-Means*; *Clustering*; KDD; PCOS; Pola Risiko

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder commonly affecting women of reproductive age and associated with reproductive and metabolic conditions. Therefore, a systematic data analysis approach is needed to identify PCOS risk patterns. This study aims to cluster PCOS patient data using the K-Means algorithm and analyze the resulting risk patterns. The data were obtained from a public Kaggle dataset consisting of 2,539 patient records, including demographic, hormonal, metabolic, reproductive, and lifestyle variables. The research methodology follows the Knowledge Discovery in Databases (KDD) process, which comprises data selection, preprocessing, transformation, data mining, evaluation, and interpretation. The K-Means algorithm was applied to group the data into three categories: low, moderate, and high risk. The results indicate that the clustering process produces groups with distinct characteristics based on data similarities. The identified patterns include characteristics such as Body Mass Index (BMI), menstrual cycle irregularities, and hormonal parameters. These findings provide an overview of PCOS patient characteristic patterns that may serve as supporting information for healthcare professionals in understanding patient data and improving public awareness of PCOS prevention and management.

Keywords: *K-Means; Clustering; KDD; PCOS; Risk Patterns*

1. PENDAHULUAN

Kondisi kesehatan tubuh berperan penting dalam menunjang aktivitas sehari-hari dan kualitas hidup seseorang. Pada perempuan, keseimbangan hormonal merupakan salah satu aspek yang berkaitan dengan fungsi reproduksi dan kondisi metabolisme. Perubahan keseimbangan tersebut dapat berhubungan dengan berbagai gangguan kesehatan, salah satunya Sindrom Ovarium Polikistik (*Polycystic Ovary Syndrome*/PCOS). Gangguan endokrin ini memiliki karakteristik yang beragam, antara lain gangguan ovulasi, peningkatan kadar androgen, dan perubahan pada ovarium. PCOS juga berkaitan dengan sejumlah kondisi metabolik, seperti resistensi insulin, obesitas, dan dislipidemia (Arfan Goni et al., 2025). Penelitian yang sama melaporkan prevalensi PCOS di Indonesia sekitar 45,7%, dengan kasus yang banyak ditemukan pada kelompok perempuan berusia 26–30 tahun serta keterkaitan dengan indeks massa tubuh (*Body Mass Index*/BMI) yang tinggi (Arfan Goni et al., 2025).

Variasi karakteristik PCOS menyebabkan pasien dapat menunjukkan kondisi reproduksi dan metabolisme yang berbeda. Walaupun obesitas, ketidakteraturan menstruasi, dan perubahan hormonal kerap dikaitkan dengan PCOS, karakteristik tersebut tidak selalu muncul dengan pola yang sama pada setiap pasien. Oleh sebab itu, pengelompokan berdasarkan kemiripan karakteristik dapat digunakan untuk mengeksplorasi keragaman pola yang terdapat dalam data pasien. Van der Ham et al. (2024), misalnya, menerapkan metode *hierarchical clustering* dan memperoleh tiga subtipe PCOS dengan perbedaan karakteristik metabolik, reproduksi, serta latar belakang pasien. Temuan tersebut memperlihatkan bahwa teknik pengelompokan dapat membantu mengidentifikasi kesamaan dan perbedaan karakteristik pasien. Meskipun demikian, diperlukan kajian lebih lanjut dengan mempertimbangkan keragaman variabel agar pola karakteristik PCOS dapat ditelaah secara lebih menyeluruh (van der Ham et al., 2024).

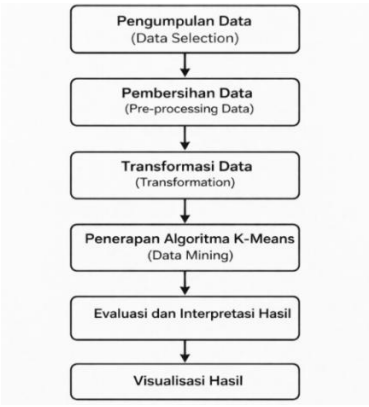
Perkembangan teknologi informasi memungkinkan data berukuran besar dianalisis untuk memperoleh pola yang tidak selalu dapat dikenali melalui pengamatan biasa. Salah satu pendekatan yang dapat dimanfaatkan adalah *data mining*, yaitu proses penggalian informasi dari kumpulan data untuk menemukan pola yang relevan (Sutiono & Wasilah, 2024). Algoritma *K-Means* merupakan salah satu metode yang dapat diterapkan untuk tujuan tersebut. Penerapannya juga telah digunakan dalam bidang kesehatan, antara lain untuk mengelompokkan puskesmas berdasarkan karakteristik kualitas kesehatan keluarga dan gizi masyarakat (Hadi Prakoso et al., 2023).

Data pasien PCOS memuat beragam atribut yang berkaitan dengan kondisi demografis, hormonal, metabolik, reproduksi, dan gejala klinis. Keragaman atribut tersebut membuka peluang untuk menerapkan *K-Means* dalam mengeksplorasi kelompok pasien berdasarkan kesamaan karakteristiknya. Penggunaan algoritma ini diharapkan dapat membantu menggambarkan variasi pola yang terdapat dalam *dataset* PCOS (Sutiono & Wasilah, 2024).

Berdasarkan kebutuhan tersebut, penelitian ini menerapkan algoritma *K-Means* pada data pasien PCOS dengan mempertimbangkan sejumlah atribut yang relevan. Fokus penelitian tidak hanya terletak pada pembentukan kelompok, tetapi juga pada penelaahan karakteristik yang membedakan setiap *cluster*. Hasil pengelompokan kemudian diinterpretasikan ke dalam kategori risiko rendah, sedang, dan tinggi berdasarkan pola atribut yang terbentuk. Kategori tersebut digunakan sebagai interpretasi hasil analisis data, bukan sebagai diagnosis maupun penilaian risiko klinis yang telah tervalidasi.

2. METODE

Pada penelitian ini digunakan Metode penelitian KDD (*Knowledge Discovery in Databases*) yang merupakan pendekatan sistematis dalam proses penggalian pengetahuan dari data. Dapat dilihat pada Gambar 1, alur KDD terdiri dari beberapa tahapan utama, yaitu: seleksi data, pembersihan (*preprocessing*), transformasi, *data mining*, serta evaluasi dan interpretasi hasil. Tahap seleksi data bertujuan untuk memilih data yang relevan dari sumber yang tersedia. Selanjutnya, *preprocessing* dilakukan untuk menghilangkan data yang tidak lengkap, duplikat, atau tidak konsisten. Transformasi data diperlukan agar data berada dalam format yang sesuai untuk dianalisis. Tahap inti adalah *data mining*, di mana algoritma seperti klasifikasi, *clustering*, atau asosiasi diterapkan untuk menemukan pola tersembunyi. Terakhir, hasil *mining* dievaluasi dan diinterpretasikan agar dapat digunakan dalam pengambilan keputusan berbasis data (Wijayati & Saputro, 2023).



Gambar 1. Alur Metode Penelitian KDD

2.1 Pengumpulan Data

Penelitian ini memanfaatkan data sekunder yang diperoleh dari Kaggle, berupa kumpulan data terstruktur yang mendukung kemudahan dalam proses pengolahan dan analisis. Dataset yang digunakan berisi data klinis dan laboratorium pasien wanita usia reproduktif untuk analisis Sindrom Ovarium Polikistik (PCOS), dalam format CSV berukuran sekitar 487 KB dengan 43 variabel dan 2539 data, yang mencakup berbagai parameter seperti kondisi fisik, hormonal, gaya hidup, serta indikator reproduksi termasuk jumlah folikel dan ukuran endometrium.

2.2 Pembersihan Data

2.2.1 Data Cleaning

Pada tahap *preprocessing*, dilakukan pembersihan data (*data cleaning*) untuk menjaga kualitas data sebelum memasuki proses analisis. Tahapan ini mencakup pemeriksaan data duplikat menggunakan fungsi *drop_duplicates()*, penanganan nilai yang hilang (*missing value*) melalui *dropna()*, serta identifikasi dan penanganan *outlier* dengan metode *Interquartile Range* (IQR) menggunakan teknik *capping*. Berdasarkan hasil pemeriksaan, ditemukan 9 data yang memiliki nilai hilang. Data tersebut kemudian dihapus agar tidak digunakan dalam analisis selanjutnya. Setelah proses pembersihan, jumlah data yang digunakan dalam penelitian tercatat sebanyak 2.539 data.

Penanganan *outlier* dilakukan dengan teknik *capping* berbasis IQR untuk membatasi nilai ekstrem tanpa menghapus keseluruhan data yang bersangkutan. Metode ini dipilih karena penghapusan *outlier* secara langsung berpotensi menghilangkan informasi yang relevan, terutama pada data klinis (Afiasari et al., 2023). Perhitungan IQR dilakukan dengan mencari selisih antara kuartil ketiga (Q3) dan kuartil pertama (Q1). Selanjutnya, batas bawah ditentukan menggunakan rumus $Q1 - 1{,}5 \times IQR$, sedangkan batas atas dihitung dengan rumus $Q3 + 1{,}5 \times IQR$. Nilai yang berada di luar rentang tersebut ditangani melalui teknik *capping* untuk mengurangi pengaruh nilai ekstrem terhadap hasil *cluster*ing. Pemeriksaan data duplikat juga menjadi bagian penting dalam proses pembersihan karena pencatatan entitas yang sama lebih dari satu kali dapat memengaruhi perhitungan jarak dan pembentukan *cluster* pada algoritma *K-Means* (Sri Astuti et al., 2021). Dengan demikian, tahap *data cleaning* ditujukan untuk menghasilkan data yang lebih konsisten sebagai dasar pelaksanaan analisis berikutnya.

2.2.2 Data Selection

Tahap ini mencakup seleksi atribut pada dataset PCOS untuk menentukan variabel yang digunakan dalam proses analisis selanjutnya. Dari 43 variabel yang tersedia, dipilih 20 atribut

yang digunakan dalam proses *cluster*ing. Pemilihan atribut didasarkan pada relevansinya terhadap karakteristik PCOS, meliputi aspek hormonal, metabolik, reproduksi, hasil pemeriksaan ultrasonografi, dan gejala klinis. Pemilihan ini dilakukan untuk mempertahankan atribut yang memiliki keterkaitan dengan kondisi PCOS sekaligus mengurangi kompleksitas data yang digunakan dalam proses analisis. Faktor obesitas, ketidakseimbangan hormon, serta adanya resistensi insulin diketahui memiliki peran yang signifikan terhadap peningkatan risiko terjadinya *Polycystic Ovary Syndrome* (PCOS) pada wanita usia reproduktif di Indonesia (Wahyuni et al., 2022).

Penyederhanaan ini bertujuan untuk mengurangi kompleksitas model, meningkatkan efisiensi komputasi, serta memastikan bahwa variabel yang digunakan benar-benar mencerminkan karakteristik penting yang berkaitan dengan risiko PCOS, karena seleksi fitur yang tepat pada algoritma *K-Means* dapat meningkatkan kualitas *cluster* dan kemudahan interpretasi hasil pengelompokan (Sekar Setyaningtyas et al., 2022). Atribut yang tidak dipilih memiliki kontribusi informasi yang rendah, redundan, atau kurang berkaitan langsung dengan kondisi klinis PCOS, sehingga jika tetap digunakan akan meningkatkan kompleksitas data dan menurunkan performa algoritma *K-Means*.

2.3 Transformasi Data

Data yang telah melalui tahap pembersihan selanjutnya dikonversi ke dalam bentuk numerik dan dinormalisasi agar setiap atribut memiliki rentang nilai yang sebanding. Proses normalisasi menggunakan metode *Min-Max Normalization* dengan mengacu pada Persamaan(1) berikut:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Keterangan:

X' : nilai atribut setelah proses normalisasi

X : nilai asli dari data

X_{min} : nilai terendah pada atribut

X_{max} : nilai tertinggi pada atribut

Normalisasi data dilakukan dengan menerapkan metode *Min-Max Normalization* untuk menyetarakan rentang nilai pada setiap atribut (Nugroho et al., 2025). Tahapan ini bertujuan untuk mencegah dominasi variabel tertentu dalam proses pengelompokan akibat perbedaan satuan maupun rentang nilai.

2.4 Penerapan Algoritma *K-Means*

a. Penentuan jumlah *cluster* (k) dilakukan dengan menetapkan banyaknya kelompok yang akan dibentuk dari data. Metode *Elbow* digunakan sebagai acuan untuk memilih jumlah *cluster* yang optimal dalam analisis *K-Means*.

b. Inisialisasi *centroid*, dilakukan dengan menentukan titik pusat awal (*centroid*) untuk setiap *cluster* secara acak. Parameter *random_state* digunakan untuk menjaga reproduktibilitas hasil, sehingga proses inisialisasi dapat menghasilkan keluaran yang sama ketika algoritma dijalankan kembali dengan parameter yang sama.

c. Menghitung jarak menggunakan *Euclidean Distance*. *Euclidean Distance* dipilih karena data yang digunakan berupa data numerik yang telah melalui proses normalisasi, sehingga jarak antarobservasi dapat dihitung berdasarkan perbedaan nilai pada setiap atribut. Jarak *Euclidean* dapat dinyatakan melalui Persamaan(2) berikut.

$$d(x,y) = \sqrt{\sum n= 1(xi - yi)2} \qquad (2)$$

Keterangan :

$d(x, y)$: jarak data x terhadap pusat *cluster* y

x_i : nilai data ke-i pada observasi

y_i : nilai titik pusat *cluster* ke-y pada observasi ke-i

n : jumlah atribut yang digunakan

d. Mengelompokkan data berdasarkan jarak terdekat, setiap data dimasukkan ke dalam *cluster* dengan *centroid* terdekat.

e. Menghitung ulang *centroid* baru yang diperoleh dari rata-rata semua data dalam masing-masing *cluster* mengacu pada Persamaan(3):

$$CI= \frac{x1+x2+x3+\cdots+xn}{\sum x} \qquad (3)$$

Keterangan:

CI : *centroid* baru

x_1 : nilai data pada *cluster* pertama

x_n : nilai data pada *cluster* ke-n

x : jumlah data

f. Iterasi hingga konvergen, dengan mengulangi proses penghitungan jarak dan pembaruan *centroid* hingga tidak ada data yang berpindah *cluster*.

g. Penggunaan *Principal Component Analysis* (PCA) merupakan teknik *unsupervised learning* yang diterapkan untuk mengurangi dimensi data dengan tetap

mempertahankan informasi penting dari data asli. PCA bekerja dengan mentransformasikan sejumlah variabel yang saling berkorelasi menjadi variabel baru yang bersifat ortogonal atau tidak berkorelasi (Rosyada & Utari, 2024). Pada penelitian ini, PCA dimanfaatkan untuk menyederhanakan dimensi data serta mendukung visualisasi hasil *cluster*ing melalui komponen utama.

h. Menilai *Silhouette Coefficient* dengan rumus *Silhouette* untuk setiap data ke-i adalah:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (4)$$

Keterangan:

$a(i)$: rata-rata jarak dalam *cluster* yang sama (*cohesion*)

$b(i)$: rata-rata jarak ke *cluster* terdekat lainnya (*separation*)

2.5 Evaluasi dan Interpretasi Hasil

Evaluasi dilakukan untuk menilai kualitas *cluster* yang dihasilkan oleh algoritma *K-Means* dengan menggunakan metrik WCSS dan *Silhouette Coefficient*. WCSS berfungsi untuk mengukur tingkat kekompakan data dalam setiap *cluster*, di mana nilai yang lebih rendah menunjukkan penyebaran data yang semakin kecil. Sementara itu, *Silhouette Coefficient* digunakan untuk menilai kedekatan data terhadap *cluster* tempatnya berada dibandingkan dengan *cluster* lainnya. Nilai *Silhouette Coefficient* yang lebih tinggi menunjukkan kualitas pengelompokan yang lebih baik. Kedua metrik tersebut digunakan sebagai acuan untuk menilai kualitas hasil *cluster*ing.

2.6 Visualisasi Hasil

Hasil akhir dianalisis menggunakan Google Colab, di mana proses *cluster*ing dilakukan dengan algoritma *K-Means*. Setelah model terbentuk, data yang telah diberi label *cluster* kemudian diekspor dan divisualisasikan melalui Streamlit. Hasil analisis disajikan dalam bentuk dashboard interaktif menggunakan Streamlit yang memuat berbagai visualisasi, meliputi distribusi data pasien, hasil pengelompokan menggunakan algoritma *K-Means*, serta karakteristik pola risiko pada masing-masing cluster. *Dashboard* ini bertujuan untuk membantu proses interpretasi hasil analisis dan mendukung pengambilan keputusan berdasarkan pola yang teridentifikasi. Hal ini sangat membantu dalam proses interpretasi hasil, karena informasi yang disajikan tidak hanya bersifat numerik, tetapi juga divisualisasikan secara grafis untuk mendukung analisis.

3. HASIL DAN PEMBAHASAN

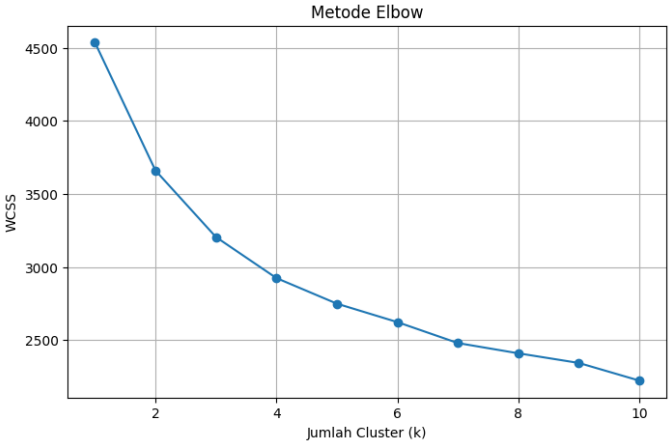
3.1 Prapemrosesan Data

Pemrosesan awal data dilakukan sebelum algoritma *K-Means* diterapkan untuk memastikan bahwa data yang digunakan telah memenuhi kebutuhan analisis *cluster*ing. *Dataset* awal yang digunakan dalam penelitian memuat 2.539 rekaman dengan 43 atribut. Pada tahap pemeriksaan kualitas data, ditemukan sembilan rekaman yang memiliki nilai kosong. Rekaman tersebut dihapus sebelum analisis dilanjutkan. Selanjutnya, dilakukan seleksi terhadap atribut yang relevan dengan karakteristik PCOS sehingga diperoleh 20 atribut untuk proses pengelompokan. Atribut terpilih meliputi aspek hormonal, metabolik, reproduksi, hasil *ultrasonografi*, dan gejala klinis. Data kemudian disiapkan dalam format numerik dan dinormalisasi menggunakan metode *Min-Max* agar perbedaan rentang nilai antara atribut tidak mendominasi perhitungan jarak. Hasil prapemrosesan selanjutnya digunakan sebagai masukan algoritma *K-Means*.

3.2 Penerapan Algoritma *K-Means*

a. Menentukan jumlah *cluster* (k):

Grafik *Elbow* pada Gambar 2 memperlihatkan penurunan nilai WCSS yang signifikan dari $k = 1$ hingga $k = 3$, kemudian penurunannya mulai melambat. Titik siku yang terbentuk pada $k = 3$ menunjukkan bahwa tiga *cluster* dipilih sebagai jumlah pengelompokan yang optimal berdasarkan metode *Elbow*.



Gambar 2. Grafik WCSS

b. Inisialisasi *centroid* secara acak:

Tabel 1. Inisialisasi *Centroid* Acak

Atribut	*Centroid* 1	*Centroid* 2	*Centroid* 3
Age (yrs)	0.314	0.471	0.549
BMI	0.455	0.763	0.514
Cycle length(days)	0.0	0.0	0.0

Atribut	*Centroid* 1	*Centroid* 2	*Centroid* 3
FSH(mIU/mL)	0.326	0.115	0.181
LH(mIU/mL)	0.075	0.016	0.100
FSH/LH	0.804	1.000	0.353
TSH (mIU/L)	0.262	0.059	0.414
AMH(ng/mL)	0.044	0.047	0.069
PRL(ng/mL)	0.541	0.245	0.547
RBS(mg/dl)	0.842	0.625	0.175
Weight gain(Y/N)	0.0	1.0	1.0
Hair growth(Y/N)	0.0	1.0	0.0
Skin darkening (Y/N)	0.0	1.0	1.0
Hair loss(Y/N)	0.0	1.0	0.0
Pimples(Y/N)	0.0	1.0	1.0
Follicle No. (L)	0.516	0.581	0.710
Follicle No. (R)	0.6	0.6	0.4
Avg. F size (L) (mm)	0.459	0.459	0.216
Avg. F size (R) (mm)	0.676	0.405	0.081
Endometrium (mm)	0.504	0.441	0.617

Tabel 1 di atas adalah hasil dari inisialisasi *centroid* yang dilakukan secara acak dengan memilih sejumlah *k* data dari *dataset X* tanpa pengulangan menggunakan fungsi *np.random.choice(). Centroid* awal ini digunakan sebagai titik pusat awal dalam proses *cluster*isasi *K-Means*.

c. Menghitung Jarak: Perhitungan jarak *Euclidean* menghasilkan matriks jarak yang menunjukkan jarak antara setiap data pasien dan masing-masing centroid pada setiap cluster.

Matriks Jarak Euclidean untuk Seluruh Data:

	Centroid 1	Centroid 2	Centroid 3
0	1.7198	2.0768	2.0254
1	2.0429	1.8115	1.6421
2	1.3466	2.7303	2.2994
3	2.1160	1.9138	2.2342
4	1.0616	2.4178	1.9540
...	...	...	...
2525	1.3954	2.8070	2.2587
2526	1.2278	2.7138	2.0910
2527	1.3188	2.6178	2.2272
2528	1.2783	2.2685	1.7564
2529	1.4112	2.7706	2.2091

Gambar 3. Sampel Jarak Setiap Data Terhadap *Centroid*

Gambar 3 menunjukkan hasil perhitungan jarak setiap data pasien terhadap masing-masing *centroid* menggunakan rumus *Euclidean*. Nilai jarak tersebut menjadi dasar penentuan *cluster*, yaitu dengan menempatkan setiap data ke dalam *cluster* yang memiliki jarak paling dekat dengan *centroid*. Proses ini dilakukan sebelum pembaruan nilai *centroid* dan pelaksanaan iterasi berikutnya dalam algoritma *K-Means*.

d. Mengelompokkan data berdasarkan jarak terdekat: Pengelompokan data dilakukan dengan membandingkan jarak *Euclidean* setiap data terhadap seluruh *centroid*. Selanjutnya, setiap data dimasukkan ke dalam *cluster* dengan nilai jarak terkecil berdasarkan hasil pencarian nilai minimum.

e. Menghitung ulang *centroid* baru: Perhitungan *centroid* baru dilakukan dengan menentukan nilai rata-rata seluruh data dalam setiap *cluster*. Pembaruan ini bertujuan untuk menyesuaikan posisi pusat *cluster* dengan distribusi data pada setiap iterasi.

f. Iterasi hingga konvergen:

```
Iterasi ke-1, Perubahan centroid = 1.6451240258383726
Iterasi ke-2, Perubahan centroid = 0.1486160809702251
Iterasi ke-3, Perubahan centroid = 0.09881479656927084
Iterasi ke-4, Perubahan centroid = 0.08325178192585365
Iterasi ke-5, Perubahan centroid = 0.06487779853658525
Iterasi ke-6, Perubahan centroid = 0.07193127358259556
Iterasi ke-7, Perubahan centroid = 0.03833387017535166
Iterasi ke-8, Perubahan centroid = 0.07351100016230469
Iterasi ke-9, Perubahan centroid = 0.04147527793535309
Iterasi ke-10, Perubahan centroid = 0.0

Proses clustering konvergen.
```

Gambar 4. Iterasi Hingga Konvergen

Pada Gambar 4 dapat dilihat iterasi awal (ke-1), perubahan *centroid* sebesar 1.6451, mencerminkan pergeseran signifikan akibat inisialisasi acak. Selanjutnya, dari iterasi ke-2 hingga ke-5, perubahan *centroid* menurun secara bertahap, menandakan *centroid* mulai menyesuaikan diri dengan distribusi data. Pada iterasi ke-6 hingga ke-9, perubahan *centroid* semakin kecil, menunjukkan stabilitas pembagian *cluster*. Pada iterasi ke-10, perubahan mencapai 0.0, menandakan *centroid* telah stabil.

3.3 Evaluasi dan Interpretasi Hasil

Tabel 2. Rekapitulasi Hasil *Clustering K-Means*

Nama *Cluster*	Interpretasi	Jumlah Data	Persentase Data
Cluster 0	Risiko Rendah	1484	58,66%
Cluster 1	Risiko Tinggi	596	23,56%
Cluster 2	Risiko Sedang	450	17,79%

Berdasarkan Tabel 2, proses *cluster*ing menghasilkan tiga kelompok yang dikategorikan sebagai risiko rendah, sedang, dan tinggi untuk mempermudah pemahaman terhadap karakteristik setiap *cluster*. Penetapan label tersebut mengacu pada perbandingan nilai *centroid* dan karakteristik variabel dalam masing-masing *cluster*, bukan berdasarkan hasil validasi atau diagnosis klinis. *Cluster* 0 mencakup 1.484 data (58,66%), *Cluster* 1 sebanyak 596 data (23,56%), sedangkan *Cluster* 2 terdiri atas 450 data (17,79%). Nilai *Silhouette Coefficient* sebesar 0,3836 mengindikasikan adanya struktur pengelompokan, meskipun masih terdapat potensi tumpang tindih (*overlap*) antar-*cluster*. Dengan demikian, label risiko dalam penelitian ini digunakan sebagai interpretasi berdasarkan pola karakteristik data dan tidak dapat dianggap sebagai tingkat risiko klinis yang telah tervalidasi.

Tabel 3. Variabel Paling Dominan

Nama Atribut	*Centroid* Risiko Tinggi	Rata-Rata *Cluster* Lain	Selisih
Hair growth(Y/N)	1,000000	0,033693	0,966307
Skin darkening (Y/N)	0,672819	0,322273	0,350546
Hair loss(Y/N)	0,711409	0,381869	0,329541
Weight gain(Y/N)	0,724832	0,475758	0,249075
Follicle No. (L)	0,578805	0,351166	0,227639
Follicle No. (R)	0.469463	0.310818	0.158645

Berdasarkan Tabel 3, sejumlah atribut memperlihatkan karakteristik yang lebih dominan pada cluster yang dikategorikan sebagai kelompok risiko tinggi. Atribut tersebut meliputi pertumbuhan rambut, kondisi kulit yang menggelap, kerontokan rambut, peningkatan berat badan, serta jumlah folikel pada ovarium kiri dan kanan. Pemberian label risiko tinggi didasarkan pada pola nilai *centroid* dan karakteristik variabel yang terbentuk dari hasil *cluster*ing, bukan berdasarkan validasi klinis. Oleh karena itu, atribut tersebut digunakan sebagai karakteristik pembeda antar*cluster* dan tidak diartikan sebagai faktor yang secara langsung memengaruhi risiko PCOS.

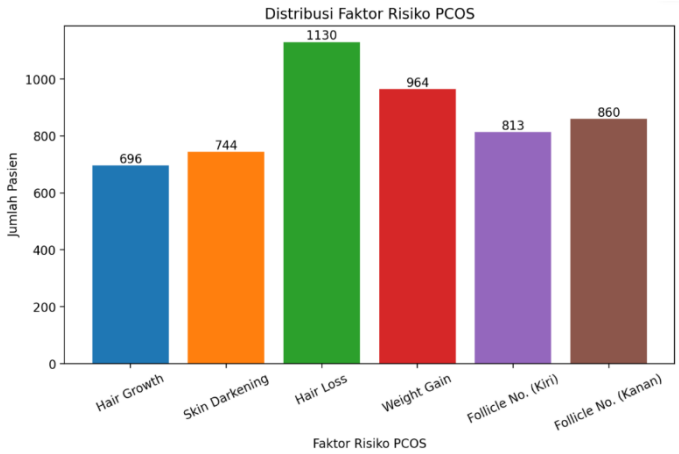
3.4 Visualisasi Hasil

Melalui *dashboard* antarmuka Streamlit, hasil klasterisasi ditampilkan dalam bentuk grafik dan ringkasan statistik. Visualisasi ini memungkinkan pengguna untuk melihat perbedaan karakteristik antar *cluster* secara lebih intuitif, terutama dalam mengidentifikasi atribut-atribut yang paling dominan pada *cluster* risiko rendah, sedang, dan tinggi. Berdasarkan Gambar 5, pada bagian kiri (*sidebar*) terdapat fitur Filter Data Pasien (K = 3) yang memungkinkan pengguna memilih *cluster* pasien (0, 1, dan 2) serta menu navigasi utama yang terdiri dari Beranda, *Data Selection, Modeling, Evaluation, Visualisasi Cluster,* dan *Insight*.

Menu ini memudahkan pengguna berpindah antar tahapan analisis sesuai metodologi *Knowledge Discovery in Databases* (KDD). Pengembangan halaman ini memadukan Python, Streamlit, Pandas, Matplotlib, dan Scikit-learn untuk mendukung penyajian hasil *cluster*ing menggunakan algoritma *K-Means*. Melalui integrasi tersebut, hasil pengelompokan data dapat ditampilkan secara interaktif dan terstruktur, sehingga memudahkan pengguna dalam melihat visualisasi serta memahami hasil analisis data pasien PCOS.



Gambar 5. Halaman Utama *Dashboard* Streamlit



Gambar 6. Diagram Batang Faktor Risiko PCOS

Diagram batang tersebut menyajikan gambaran jumlah pasien PCOS berdasarkan masing-masing faktor risiko yang diamati. Sumbu horizontal menunjukkan jenis faktor risiko PCOS, yaitu *hair growth, skin darkening, hair loss, weight gain,* dan jumlah folikel tinggi, sedangkan sumbu vertikal merepresentasikan jumlah pasien yang mengalami kondisi tersebut. Perbedaan tinggi batang pada grafik menunjukkan bahwa gangguan kerontokan rambut dan peningkatan berat badan merupakan faktor risiko PCOS yang paling dominan dalam dataset penelitian ini, sehingga memiliki peran penting dalam analisis karakteristik klinis pasien PCOS.

4. PENUTUP

Simpulan dan Saran

Berdasarkan hasil analisis, metode *Elbow* digunakan untuk menentukan jumlah *cluster* optimal, yaitu tiga kelompok pasien PCOS dengan karakteristik yang serupa. Ketiga kelompok tersebut kemudian diinterpretasikan sebagai kategori risiko rendah, sedang, dan tinggi. Pada kelompok risiko tinggi, karakteristik yang paling menonjol meliputi pertumbuhan rambut berlebih, penggelapan kulit, kerontokan rambut, peningkatan berat badan, serta jumlah folikel pada ovarium kiri dan kanan. Nilai *Silhouette* sebesar 0,3836 menunjukkan adanya struktur pengelompokan, meskipun masih terdapat kemungkinan tumpang tindih antar*cluster*. Hasil *cluster*ing dalam penelitian ini digunakan untuk menggambarkan pola karakteristik data dan tidak dimaksudkan sebagai penentuan diagnosis atau risiko klinis secara langsung. Penelitian ini memiliki keterbatasan pada penggunaan dataset sekunder dari Kaggle yang representasinya terhadap populasi tertentu belum dapat dipastikan. Selain itu, penelitian menggunakan algoritma *K-Means* dengan inisialisasi *centroid* dan dua metrik evaluasi, yaitu WCSS dan *Silhouette*. Penelitian selanjutnya dapat menggunakan dataset yang lebih beragam dan representatif, membandingkan *K-Means* dengan metode *cluster*ing lainnya, serta melakukan pengujian dengan lebih banyak metrik dan beberapa inisialisasi *centroid* untuk mengetahui konsistensi hasil *cluster*ing.

DAFTAR PUSTAKA

Afiasari, N., Suarna, N., & Rahaningsi, N. (2023). Implementasi Data Mining Transaksi Penjualan Menggunakan Algoritma Clustering dengan Metode *K-Means*. *Jurnal SAINTEKOM*, *13*(1), 100–110. https://doi.org/10.33020/saintekom.v13i1.402

Arfan Goni, P., Rani Oktapiani, D., Fadya Nur Shaumi, N., Fadlilah, F., Rama Noufalwafiq, M., & Ariyanto, J. (2025). Aspek Klinis, Hormonal dan Psikososial Sindrom Ovarium pada Remaja : Tinjauan Pustaka. *Agustus*, *4*(4), 43–54.

Hadi Prakoso, B., Rachmawati, E., Rachmatta Putro Mudiono, D., Vestine, V., & Eko Julianto Suyoso, G. (2023). Klasterisasi Puskesmas dengan *K-Means* Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat. *Jurnal Buana Informatika*, *14*(1), 60–68.

Nugroho, N., Wijiyanto, W., & Pradana, A. I. (2025). Analisis Persebaran Izin Usaha di Kabupaten Sragen Menggunakan Metode *K-Means* Clustering. *Jurnal Pendidikan Dan Teknologi Indonesia*, *5*(1), 143–153. https://doi.org/10.52436/1.jpti.608

Rosyada, I. A., & Utari, D. T. (2024). Penerapan Principal Component Analysis untuk Reduksi Variabel pada Algoritma *K-Means* Clustering. *Jambura J. Probab. Stat*, *5*(1), 6–13. https://doi.org/10.34312/jjps.v5i1.18733

Sekar Setyaningtyas, Indarmawan Nugroho, B., & Arif, Z. (2022). TINJAUAN PUSTAKA SISTEMATIS: PENERAPAN DATA MINING TEKNIK CLUSTERING ALGORITMA *K-MEANS*. *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, *10*(2), 52–61. https://doi.org/10.21063/jtif.2022.v10.2.52-61

Sutiono, & Wasilah. (2024). PERBANDINGAN KINERJA ALGORITMA *K-MEANS* DAN K-MEDOIDS UNTUK MENGELOMPOKAN DATA SISWA PENERIMA BOSDA DI SMKN1 KATIBUNG. *Jurnal Teknologi Informasi Magister Darmajaya*, *1*(1). https://journal.darmajaya.ac.id/index.php/jtim/

van der Ham, K., Moolhuijsen, L. M. E., Brewer, K., Sisk, R., Dunaif, A., Laven, J. S. E., Louwers, Y. V., & Visser, J. A. (2024). Clustering Identifies Subtypes With Different Phenotypic Characteristics in Women With Polycystic Ovary Syndrome. *Journal of Clinical Endocrinology and Metabolism*, *109*(12), 3096–3107. https://doi.org/10.1210/clinem/dgae298

Wahyuni, A., Supriyatiningsih, Kusumawati, W., Sedah Kirana, K., & Mayayustika, C. D. (2022). Family history of PCOS, obesity, low fiber diet, and low physical activity increase the risk of PCOS. *Jurnal Kedokteran Dan Kesehatan Indonesia*. https://doi.org/10.20885/jkki.vol13.iss1.art8

Wijayati, R., & Saputro, D. R. S. (2023). *Clustering Data Numerik dan Kategorik Menggunakan Algoritme K-Prototype. PRISMA, Prosiding Seminar Nasional Matematika 6*. 702–706. https://journal.unnes.ac.id/sju/index.php/prisma/