Information System Journal (INFOS)
Vol. 8, No. 1, Mei 2025, pp. 51-59
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2025v8i01.2081

IDENTIFIKASI FAKTOR UTAMA PENYEBAB SINDROM OVARIUM POLIKISTIK (PCOS) MENGGUNAKAN ALGORITMA C4.5

Aziza Rahma[1)], **Anna Ariyantina**[2)], **Hidayanti Murtina**[3)]
[1) 2) 3)] *Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika*

email: 19210633@bsi.ac.id[1)], 19210787@bsi.ac.id[2)] , hidayanti.hym@bsi.ac.id[3)]

INFO ARTIKEL

Riwayat Artikel:

Diterima Mei, 2025
Revisi Mei, 2025
Terbit Mei, 2025

ABSTRAK

Sindrom Ovarium Polikistik (*SPOK*) merupakan kondisi hormonal umum pada wanita usia reproduktif, ditandai dengan disfungsi *ovarium*, kadar *androgen* tinggi, dan resistensi *insulin*. Menurut *WHO*, 6–13% wanita mengalami *PCOS*, dan hingga 70% tidak terdiagnosis. Penelitian ini bertujuan untuk memprediksi faktor utama penyebab terjadinya *PCOS* dengan menggunakan *algoritma C4.5* berdasarkan atribut data klinis, seperti *Body Mass Index* (*BMI*), *Menstrual Irregularity*, *Testosterone Level* (*ng/dL*), dan *Antral Follicle Count*. Dapat disimpulkan bahwa atribut *Menstrual Irregularity* merupakan faktor paling dominan diikuti oleh *BMI*, *Testosterone Level*(*ng/dL*), dan *Antral Follicle Count*. Model yang dikembangkan mampu mencapai akurasi sebesar 83%, dengan *precision* mencapai 94% dan *recall* sebesar 78%. Hasil ini mengindikasikan bahwa model tersebut memiliki kemampuan yang sangat baik dalam mengenali kasus positif *PCOS* dengan tingkat kesalahan yang minim. Perbandingan antara hasil perhitungan manual menggunakan *Excel* dan proses otomatis melalui *RapidMiner* menghasilkan struktur pohon yang sama, sehingga menegaskan kredibilitas dan konsistensi metode yang digunakan.

Kata Kunci :
Algoritma C4.5; Ketidakteraturan Menstruasi; Sindrom Ovarium Polikistik

ABSTRACT

Polycystic ovary syndrome (PCOS) is a common hormonal disorder in women of reproductive age, characterized by ovarian dysfunction, high androgen levels and insulin resistance. According to WHO, 6-13% of women have PCOS, and up to 70% are undiagnosed. This study aims to predict the main factors causing PCOS by using the C4.5 algorithm based on clinical data attributes, such as BMI, Menstrual Irregularity, Testosterone Level(ng/dL), and Antral Follicle Count. It can be concluded that the Menstrual Irregularity attribute is the most dominant factor followed by BMI, Testosterone Level (ng/dL), and Antral Follicle Count. The developed model achieved 83% accuracy, 94% precision, and 78% recall, showing strong capability in identifying positive PCOS cases with minimal error rate. Comparison between the results of manual calculation using Excel and the automatic process through RapidMiner resulted in the same tree structure, thus confirming the credibility and consistency of the method used.

***Keywords*:**
Algorithm C4.5; Menstrual Irregularity; Polycystic Ovary Syndrome

Penulis Korespondensi:

Aziza Rahma
Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika

Email:
19210633@bsi.ac.id

1. PENDAHULUAN

Sindrom Ovarium Polikistik (*SPOK*), yang lebih dikenal dengan istilah *Polycystic Ovary Syndrome* (*PCOS*), adalah suatu kondisi hormonal yang umum dialami oleh perempuan pada masa reproduksinya yang sering dihubungkan dengan pertumbuhan *ovarium* yang tidak normal, kadar *androgen* yang tinggi, serta

resistensi insulin, semua ini merupakan risiko potensial untuk penyakit jantung [1]. *Polycystic Ovary Syndrome* (*PCOS*) disebut juga sebagai *Sindrom Stein-Leventhal* yang diambil dari nama dua dokter yang pertama kali mendeskripsikannya pada tahun 1935 [2]. Menurut *World Health Organization*, *PCOS* berdampak pada sekitar 6–13% wanita dalam usia reproduktif, dan hingga 70% kasusnya tidak terdiagnosis [3].

Penelitian [4] mengungkapkan bahwa, dari 194 peserta yang tinggal di Kanada, dengan 93% di antaranya berada di Alberta, ditemukan bahwa diagnosis *PCOS* biasanya terjadi sekitar 4,3 tahun setelah mereka mulai menyadari gejala pertama. Selain itu, 57% responden harus berkonsultasi dengan lebih dari satu penyedia layanan kesehatan primer sebelum mendapatkan diagnosis tersebut [4]. Hal yang memprihatinkan, setengah dari responden (53%) melaporkan tidak mendapatkan rujukan untuk perawatan lanjutan ke spesialis [4]. Lebih dari itu, 70% responden tidak mendapatkan informasi mengenai risiko kesehatan jangka panjang [4]. Seperti penumpukan lemak di hati, kesulitan dalam mengolah glukosa, ketidakseimbangan lipid, tekanan darah tinggi, diabetes tipe II, dan masalah pada sistem reproduksi [5].

Gejala umum penderita *PCOS* mencakup kesulitan untuk hamil, keberadaan *ovarium* yang memiliki banyak kista, peningkatan hormon *androgen*, resistensi terhadap *insulin* atau kadar *insulin* yang tinggi, serta perlambatan dalam pertumbuhan *folikel* di *ovarium* [6]. Namun, faktor risiko utama *PCOS* pada wanita ialah riwayat keluarga yang mencapai 43% dan obesitas sebesar 34%, sementara faktor risiko lainnya teridentifikasi di bawah 30% [7]. Di Indonesia, angka ketidaksuburan di antara wanita berusia 30-34 tahun adalah 15%, lalu meningkat menjadi 30% pada usia 35-39 tahun, dan melambung sampai 55% pada usia 40-44 tahun [8]. Berdasarkan penelitian yang dilakukan oleh Perhimpunan Rumah Sakit Seluruh Indonesia (PERSI) di Jakarta, *prevalensi infertilitas* pada pria tercatat sebanyak 36%, sedangkan pada wanita mencapai 64% [8]

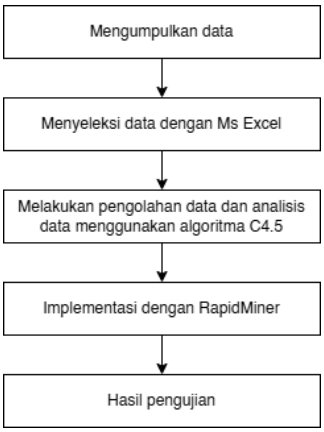
Pada penelitian sebelumnya yang pernah dilakukan pada studi kasus seleksi fitur algoritma genetika klasifikasi data rekam medis *Polycystic Ovary Syndrome* (*PCOS*) menggunakan *Support Vector Machine* (*SVM*) menghasilkan pemilihan variabel dari algoritma genetika, terdapat 475% variabel terpilih atau sembilan belas variabel yang signifikan dengan akurasi = 82.46%, sensitivitas = 60.91%, dan spesifisitas = 97.25% [9]. Pada studi kasus lainnya analisis algoritma *Regresi Logistik Biner* pada penyakit *Polycystic Ovary Syndrome* (*PCOS*) menunjukkan dari delapan variabel bebas yang dievaluasi sebagai faktor-faktor yang mempengaruhi penyakit *Polycystic Ovary Syndrome* (*PCOS*), tiga yang sangat penting adalah indeks massa tubuh kategori obesitas (*X_2~3*), diabetes (*X_3*), dan kadar hormon *anti-müllerian* (*X_7*) [10]. Kemudian penelitian menggunakan algoritma *C4.5* terbukti efektif dalam menghasilkan aturan-aturan dalam penanaman cabai dan menghasilkan akurasi sebesar 84.03% [11]. Selain itu, penelitian yang menggunakan algoritma *C4.5* juga berhasil menemukan faktor penting dalam pengambilan keputusan UKT [12].

Penelitian ini bertujuan untuk memprediksi faktor utama penyebab terjadinya *PCOS* dengan menggunakan algoritma *C4.5*. Metode algoritma *C4.5* dipilih karena kemampuannya yang luar biasa dalam mengelola *dataset* yang rumit dan mampu menghasilkan aturan keputusan yang jelas. Penelitian ini memiliki batasan, yaitu hanya akan mempertimbangkan penyebab *PCOS* berdasarkan atribut data klinis, seperti, *Body Mass Index* (*BMI*), *Menstrual Irregularity*, *Testosterone Level(ng/dL)*, dan *Antral Follicle Count*. Penggunaan algoritma *C4.5* tidak hanya menghasilkan nilai akurasi, tetapi juga menghasilkan pohon faktor yang nantinya akan memudahkan dalam mendeteksi dini penyebab *PCOS*. Sehingga, penelitian ini tidak hanya bertujuan untuk mendeteksi faktor utama penyebab *PCOS*, tetapi juga meningkatkan pemahaman masyarakat mengenai *PCOS*.

2. METODOLOGI PENELITIAN

Penelitian ini termasuk dalam jenis penelitian kuantitatif yang memanfaatkan *algoritma C4.5* dengan tujuan untuk menentukan faktor-faktor utama yang berkontribusi sebagai penyebab *Sindrom Ovarium Polikistik* menggunakan *algoritma C4.5* dalam klasifikasi data. Sumber data dalam studi ini diperoleh dari *Kaggle*, yang bersifat data publik atau terbuka. Atribut data yang dikumpulkan seperti, *Body Mass Index* (*BMI*), *Menstrual Irregularity*, *Testosterone Level(ng/dL)*, dan *Antral Follicle Count*. Gambar 1., menunjukkan tahapan-tahapan penelitian, yang menunjukkan alur kerja yang dilakukan.

Information System Journal (INFOS)
Vol. 8, No. 1, Mei 2025, pp. 51-59
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2025v8i01.2081



Gambar 1. Tahapan Penelitian [14].

Pengumpulan data dilakukan berdasarkan temuan data, pada data *open source* di salah satu website yaitu Kaggle. Proses ini mencakup akses ke sistem atau basis data terbuka di situs *website* yang menyimpan informasi terkait gejala awal dari Sindrom Ovarium Polikistik. Setelah data temuan terkumpul, dilakukan pra pemrosesan data. Proses ini mencakup; menghilangkan informasi yang salah, ganda, atau tidak berkaitan serta mengidentifikasi fitur yang paling relevan atau memengaruhi klasifikasi anggota satuan. Selanjutnya, menangani informasi yang tidak utuh atau nilai yang tidak biasa dengan cara memperbaiki atau menghapus data yang hilang serta mengendalikan informasi yang berada di luar batas normal. Tahap berikutnya melibatkan pemilihan fitur melalui berbagai cara seperti analisis hubungan, analisis data, atau pemilihan yang didasarkan pada model untuk menemukan atribut yang paling krusial dan memberikan informasi yang relevan dalam proses pengelompokan. Setelah memilih fitur-fitur yang signifikan dan mengatur data dengan cermat, barulah menerapkan algoritma C4.5 untuk membuat pohon keputusan berdasarkan atribut yang telah ditentukan.

Berikut merupakan tahapan dalam studi ini:

a. **Pengelompokan Data**
   Untuk mempermudah analisis, data diorganisir dengan memanfaatkan *Microsoft Excel*. Proses ini dilakukan dengan menata data sesuai dengan gejala atribut yang relevan.

b. **Mengolah dan menganalisis data**
   Setelah dilakukan pengelompokan, data akan diproses serta dianalisis menggunakan *algoritma C4.5*. Selanjutnya, informasi mengenai pasien dari rekam medis yang tersedia dalam data *open source kaggle* akan diklasifikasikan dan diproses untuk menciptakan pohon keputusan.

c. **Penggunaan *RapidMiner***
   Data yang sudah diproses, diuji dengan menggunakan aplikasi *RapidMiner* untuk memeriksa hasil akhirnya, apakah struktur pohon sudah sesuai dengan perhitungan manual atau tidak.

d. **Hasil Penelitian**
   Masa evaluasi faktor-faktor utama gangguan *Sindrom Ovarium Polikistik* berdasarkan, *Body Mass Index* (*BMI*), *Menstrual Irregularity*, *Testosterone Level(ng/dL)*, dan *Antral Follicle Count* dari pengolahan klasifikasi data sebelumnya dengan menggunakan *Microsoft Excel* dan aplikasi *RapidMiner* dalam bentuk *decision tree* dan aturan yang mencerminkan langkah-langkah dalam menemukan masalah serta tujuan penelitian.

Algoritma C4.5, yang dikembangkan oleh *John Ross Quinlan*, merupakan peningkatan dari *algoritma ID3.* Berbeda dengan *ID3* yang mengandalkan *Information Gain*, *C4.5* memakai *Gain Ratio* untuk mencegah adanya bias dalam memilih atribut pemisah terbaik [13]. Algoritma ini diperkenalkan pada tahun 1993 dan bisa diaplikasikan untuk data yang bersifat kategorikal maupun numerik (kontinu). Dalam hal data numerik, algoritma ini menciptakan batasan nilai (*thresholds*) dan membagi data menjadi beberapa interval guna menghasilkan nilai-nilai kategoris. Algoritma ini juga mampu menangani data dengan atribut yang hilang, yang diberi simbol "? " dan tidak dimasukkan dalam perhitungan *information gain* dan *entropy* [13].

Sangat penting untuk memeriksa tahapan perhitungan *algoritma C4.5* secara manual jika ingin memahami cara kerjanya secara menyeluruh. Metode ini tidak hanya menjelaskan konsep dasar proses pembentukan pohon keputusan, tetapi juga menjelaskan bagaimana algoritma ini menemukan atribut terbaik untuk digunakan sebagai pemisah data. Oleh karena itu, langkah-langkah perhitungan manual *algoritma C4.5* akan dijelaskan secara menyeluruh di bagian berikut, mulai dari menghitung nilai *entropy* dan keuntungan informasi hingga menemukan rasio keuntungan yang menjadi dasar untuk memilih atribut terbaik. Penjelasan ini diharapkan mampu memberi pemahaman yang jelas mengenai tahapan pengambilan keputusan yang dilakukan oleh *algoritma C4.5* dalam konteks data yang digunakan.

Berikut uraian mengenai langkah-langkah perhitungan manual algoritma C4.5.

1. **Menyusun data dan memilih atribut tujuan**
   Langkah pertama adalah menyiapkan data dalam bentuk tabel, yang mengandung satu atribut target (kelas atau label) dan beberapa atribut. Misalnya, apakah cuaca, suhu, kelembapan, dan angin memengaruhi pilihan seseorang untuk bermain golf.
2. **Menghitung tingkat ketidakpastian (e*ntropy*) dari total data**
   Untuk mengukur ketidakpastian dalam kumpulan data, *entropy* digunakan. Di sini, kita menghitung seberapa "bercampur" data berdasarkan label target. Jika semua data berada dalam satu kelas, maka tidak ada *entropy*, yang berarti tidak ada ketidakpastian.

   Rumus *Entropy*:

   $$\text{Entropy}\ (S) = \sum_{i=1}^{n} - P_i \log_2(P_i) \quad (1)$$

   Keterangan:
   S : Himpunan data
   P_i : Proporsi data untuk kelas ke-i
   n : Jumlah kelas dalam data

   Misal: 10 data, 6 “Ya” , 4 “Tidak”, maka:

   $$\text{Entropy}\ (S) = -\left(\frac{6}{10} log_2 \frac{6}{10} + \frac{4}{10} log\ 2\frac{4}{10}\right) = -(0.6\ log_2\ 0.6 + 0.4\ log_2\ 0\ .4) = 0.971 \quad (2)$$

3. **Menghitung *entropy* masing-masing atribut**
   Setiap atribut dikelompokkan berdasarkan nilainya. Misalnya, "cerah", "mendung", atau "hujan" adalah atribut cuaca. Selanjutnya, berdasarkan target, tingkat ketidakpastian dalam masing-masing kelompok dihitung. Ini menunjukkan seberapa baik atribut membagi data.
4. **Menghitung nilai informasi yang didapat (*Information Gain*)**
   Setelah mengetahui seberapa "campur" data dalam setiap nilai atribut, kita dapat menghitung seberapa besar penurunan ketidakpastian yang terjadi jika kita membagi data dengan atribut tersebut. Semakin besar penurunan ketidakpastian yang dihasilkan oleh atribut, semakin baik atribut itu untuk digunakan sebagai pemisah (*node*).

   Rumus *Info Gain*:

   $$Info\ Gain\ (\text{S}, \text{A}) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (3)$$

   Keterangan:
   S : Himpunan kasus
   A : Atribut
   n : Jumlah partisi Atribut A
   $|S_i|$: Jumlah kasus pada partisi ke-i
   $|S|$: Jumlah kasus dalam S
5. **Memilih atribut terbaik untuk membuat *node* pohon**
   Kita membuat *node* utama (akar) pohon keputusan dengan memilih atribut dengan rasio keuntungan tertinggi. Nilai-nilai atribut akan sesuai dengan cabang-cabang *node* ini. Proses untuk menghitung akar dari pohon dimulai dengan memilih atribut dengan nilai gain tertinggi. Langkah awal adalah melakukan perhitungan nilai *entropy* dari atribut yang tersedia.

Information System Journal (INFOS)
Vol. 8, No. 1, Mei 2025, pp. 51-59
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2025v8i01.2081

Beberapa karakteristik utama dari algoritma C4.5 adalah:
a. Menyediakan data latih (*training*).
b. Menetapkan akar dari pohon.
c. Menghitung nilai gain untuk setiap atribut [15].

6. **Mengulangi proses untuk setiap cabang**
Proses dari langkah 2 sampai 6 diulang untuk setiap cabang yang terbentuk, tetapi hanya untuk subset data yang sesuai dengan cabang tersebut. Pohon menurun hingga:
a. Cabang itu memiliki kelas yang sama, atau
b. Tidak ada fitur tambahan yang dapat digunakan

7. **Membuat node daun (*Leaf Node*)**
Jika tidak ada lagi data yang dapat dibagi atau jika semua data sudah homogen (semua kelas sama), cabang itu diputuskan dan kita buat daun pohon untuk menunjukkan hasil klasifikasi.

3. HASIL DAN PEMBAHASAN

Pemrosesan data ini menggunakan *algoritma C4.5* untuk menemukan faktor utama yang berdampak pada diagnosis *PCOS*. *Algoritma C4.5* diharapkan dapat menciptakan pohon keputusan yang dimulai dari bagian atas (*root*) dan terus berkembang ke arah bawah (*end*). Dalam struktur ini, atribut yang terdapat di tingkat atas disebut sebagai *node* atau *root*, yang mewakili akar atau atribut, sementara tingkat bawah dikenal sebagai daun, yang mewakili kelas.

Berikut adalah langkah-langkah untuk menerapkan *algoritma C4.5*.

1. **Menyiapkan Data**

Data yang diterapkan dalam studi ini dikumpulkan menggunakan metode dokumentasi. Proses pengumpulan data dilakukan dengan cara mengunduh berkas *CSV* yang ada di situs *https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset*. *File CSV* tersebut memuat informasi yang mencakup lima atribut yang mewakili diagnosis awal penderita PCOS.

Dataset ini berisi 500 data yang nantinya akan dilakukan perhitungan manual guna mengidentifikasi dari keempat atribut tersebut, atribut mana yang menjadi urutan paling utama dalam mendiagnosis *PCOS* dini. Teknik pengumpulan data menggunakan data sekunder. Tabel 1 merujuk pada contoh kumpulan data mengenai *PCOS* yang diperoleh dari situs *web Kaggle*.

Tabel 1. Cuplikan Data.

No	*BMI*	*Menstrual Irregularity*	*Testosterone Level (ng/dL)*	*Antral Follicle Count*	*PCOS Diagnosis*
1	31.3	*Normal*	42.7	25	*Negative*
2	24	*Normal*	30.7	14	*Negative*
3	19	*Abnormal*	70.4	23	*Negative*
4	28	*Abnormal*	77	20	*Positive*
5	21.6	*Normal*	76.8	9	*Negative*
…	…	…	…	…	…
496	31.1	*Abnormal*	83.1	25	*Positive*
497	29.7	*Abnormal*	98.7	14	*Positive*
498	26.6	*Abnormal*	42.2	21	*Positive*
499	22.1	*Abnormal*	59.8	8	*Negative*
500	19.3	*Normal*	28.4	6	*Negative*

2. **Menghitung total nilai *entropy* dari keseluruhan data berdasarkan kategori label kelas**

Pada langkah ini, dihitung nilai *entropy* keseluruhan dari semua data (500 data) untuk menilai seberapa besar ketidakpastian (*impurity*) terhadap label kelas yang ada. *Entropy* diterapkan dalam algoritma pohon keputusan (seperti *C4.5*) untuk menilai tingkat homogenitas sebuah dataset dalam konteks klasifikasi. Tabel 2 merujuk pada perhitungan total *entropy* berdasarkan 500 data.

Tabel 2. *Entropy* Total Seluruh Data.

Jumlah	Label (*PCOS Diagnosis*)		*Entropy*
	Positive	*Negative*	
500	199	301	0,969769

$$Entropy \text{ total } = \sum_{i=1}^{n} - P_j \log_2(P_j) \quad (4)$$

$$Entropy \text{ total } = \left(-\left(\frac{Positive}{Jumlah}\right) \times log_2\left(\frac{Positive}{Jumlah}\right)\right) + \left(-\left(\frac{Negative}{Jumlah}\right) \times log_2\left(\frac{Negative}{Jumlah}\right)\right) \quad (5)$$

$$Entropy \text{ total } = \left(-\left(\frac{199}{500}\right) \times log_2\left(\frac{199}{500}\right)\right) + \left(-\left(\frac{301}{500}\right) \times log_2\left(\frac{301}{500}\right)\right) = 0{,}969768 \quad (6)$$

3. Menghitung nilai *entropy* dan *info gain* untuk masing-masing atribut

Setelah melakukan perhitungan total *entropy* dari seluruh data, tahap berikutnya adalah menghitung *entropy* dan informasi yang diperoleh untuk setiap atribut yang ada. Tujuan dari langkah ini adalah untuk memahami seberapa besar peran masing-masing atribut dalam mengurangi ketidakpastian terkait label kelas. Proses ini dilakukan dengan cara mengelompokkan data berdasarkan nilai unik dari setiap atribut, kemudian menghitung *entropy* lokal untuk setiap bagian tersebut. Tabel 3 merujuk pada perhitungan *entropy* dan *information gain* pada atribut diskrit yaitu *Menstrual Irregularity.*

Tabel 3. *Entropy* dan *Information Gain* pada Atribut Diskrit.

Atribut		Jumlah	PCOS Diagnosis		*Entropy*	*Info Gain*
			Positive	*Negative*		
Menstrual Irregularity	*Abnormal*	332	199	133	0,971302	0,324824
	Normal	168	0	168	0	

Pada atribut *menstrual irregularity* di mana mengandung dua nilai, yaitu *abnormal* dan *normal*. Berdasarkan perhitungan Tabel 3., didapat kesimpulan bahwa *menstrual irregularity* yang memiliki nilai *normal* memiliki *PCOS Diagnosis positive*, yang mana *menstrual irregularity* juga sebagai *root* atau akar.

Khusus atribut numerik, sebelum dilakukan perhitungan *entropy* dan *gain*, terlebih dahulu dilakukan pemecahan dengan mengambil nilai terbaik dengan *gain* terbesar (bisa dengan nilai *mean* atau *median*). Tabel 4., memulai dengan menghitung *median* terlebih dahulu, sebelum melakukan perhitungan nilai *entropy* dan *gain* untuk setiap atribut numerik.

Tabel 4. *Median* untuk Atribut Numerik.

Median		
BMI	***Testosterone Level(ng/dL)***	***Antral Follicle Count***
27,7	63,2	19

Perhitungan dilakukan untuk menentukan nilai *entropy* dan *information gain* untuk setiap atribut numerik, seperti *BMI*, *Testosterone Level*(ng/dL), dan *Antral Follicle Count*. Selanjutnya, pilih nilai *information gain* yang tertinggi untuk merepresentasikan kinerja atribut dalam mengelompokkan data. Atribut yang memiliki *gain* maksimum akan dipilih sebagai simpul berikutnya dalam struktur pohon keputusan, mengikuti atribut utama (*Menstrual Irregularity*) yang telah ditetapkan sebagai simpul akar. Tabel 5 menunjukkan perhitungan *entropy* dan *gain* untuk setiap atribut yang memiliki nilai numerik.

Tabel 5. *Entropy* dan *Information Gain* untuk Atribut Numerik.

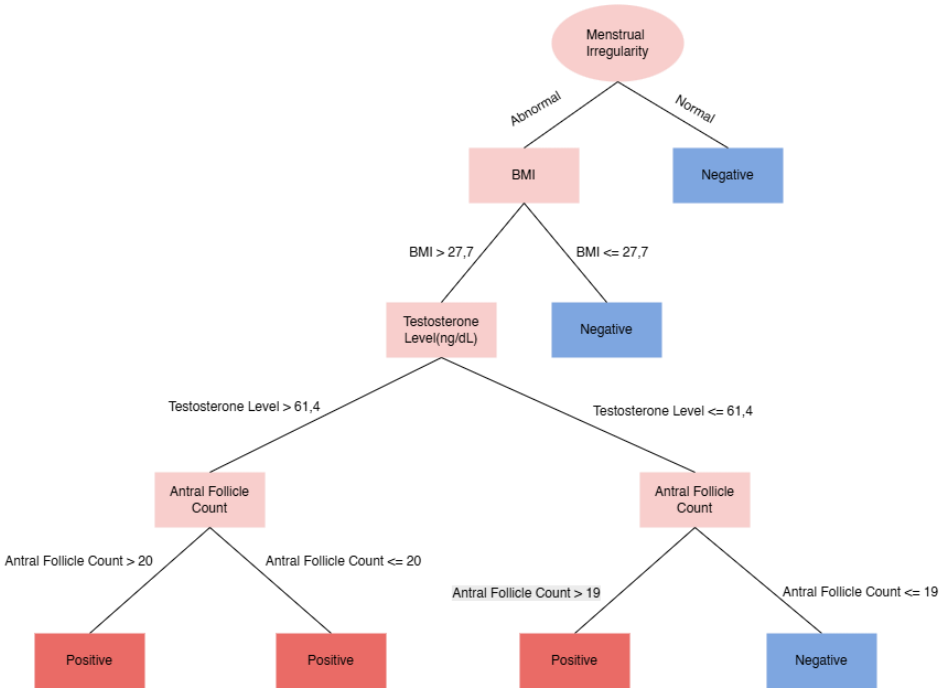
		PCOS Diagnosis			*Entropy*	*Information Gain*	*Gain Max*
		Jumlah	***Positive***	***Negative***			
BMI	<= 27,7	253	52	201	0,732858	1,079961	1,079961
	> 27,7	247	147	100	0,973722		
Testosterone Level(ng/dL)	<= 63,2	250	83	167	0,916957	1,009419	
	> 63,2	250	116	134	0,996257		
Antral Follicle Count	<= 19	264	92	172	0,932708	0,946336	
	> 19	236	107	129	0,993722		

Berdasarkan perhitungan Tabel 5., dapat disimpulkan bahwa *gain max* nya adalah *BMI* dan untuk *BMI* <= 27.7, *PCOS Diagnosis negative*. Namun untuk *BMI* > 27.7 masih membingungkan karena nilai *positive* nya masih lebih besar, maka dari itu dilakukan teknik *pruning*, yaitu dengan menyederhanakan struktur pohon dengan menghilangkan data yang nilai *BMI* <= 27.7. Lakukan perhitungan berulang sama seperti sebelumnya sampai tidak ada atribut yang tersisa untuk dibagi.

Information System Journal (INFOS)
Vol. 8, No. 1, Mei 2025, pp. 51-59
E-ISSN : 2655-142X , P-ISSN : 2655-190X, DOI:https://doi.org/10.24076/infosjournal.2025v8i01.2081

4. Hasil pohon Keputusan

Bagian ini menyampaikan hasil akhir dari struktur pohon keputusan yang dibuat melalui perhitungan manual dengan menggunakan *algoritma C4.5*. Struktur dari pohon ini dibangun dengan mempertimbangkan analisis *entropy* dan *information gain* dari setiap atribut yang terdapat dalam kumpulan data. Proses pemilihan atribut dilakukan secara bertahap, dimulai dari atribut yang memiliki nilai *information gain* tertinggi yang berfungsi sebagai simpul akar, kemudian dilanjutkan dengan pemisahan cabang berdasarkan nilai atribut yang relevan. Setiap cabang menunjukkan langkah-langkah keputusan yang berujung pada penentuan akhir, yaitu apakah individu tersebut terdiagnosis mengalami *PCOS* atau tidak. Gambar 2 menggambarkan *output* dari pohon keputusan yang diperoleh melalui perhitungan manual dengan menggunakan *Microsoft Excel*.



Gambar 2. Pohon Keputusan Perhitungan Menggunakan *Microsoft Excel*.

5. Pengujian data

Pengujian data dilakukan untuk memastikan apakah label kelas (*PCOS Diagnosis*) memenuhi kriteria dari perhitungan yang telah dilakukan. Uji coba ini didasarkan pada keputusan yang dihasilkan dari proses perhitungan sebelumnya. Tujuannya adalah untuk mengevaluasi apakah label *PCOS Diagnosis positive* atau *negative* sudah sesuai dengan ketentuan yang berlaku. Pengujian data ini sebanyak 150 data. Tabel 6., mengacu pada contoh hasil dari pengujian data yang telah dilaksanakan.

Tabel 6. Data Uji.

No	*BMI*	*Menstrual Irregularity*	*Testosterone Level(ng/dL)*	*Antral Follicle Count*	*PCOS Diagnosis*	*Prediction*
1	29,4	*Abnormal*	88,6	5	*Negative*	*Positive*
2	34,9	*Abnormal*	20	15	*Negative*	*Negative*
3	28,1	*Abnormal*	56,7	12	*Positive*	*Negative*
…		…		…		…
148	27,7	*Abnormal*	79,9	23	*Positive*	*Negative*
149	24,3	*Normal*	73,4	29	*Negative*	*Negative*
150	25,9	*Normal*	70,8	13	*Negative*	*Negative*

6. Menghitung *performance*

Berdasarkan hasil pengujian sebelumnya, dilakukan perhitungan *performance* guna untuk menilai seberapa efektif model beroperasi, memahami tipe kesalahan dalam prediksi, serta memastikan bahwa hasil yang diperoleh sudah sesuai atau belum. Namun sebelum itu dilakukan perhitungan *confusion matrix* guna untuk mengetahui jumlah prediksi benar atau salah, kemudian barulah dihitung *performance vector* seperti *accuracy*, *precision*, dan *recall*. Tabel 7., merujuk pada perhitungan *confusion matrix* terlebih dahulu.

Tabel 7. *Confusion Matrix.*

Aktual	*Diagnosis Prediction*	
	Positive	***Negative***
Positive	47	22
Negative	5	76

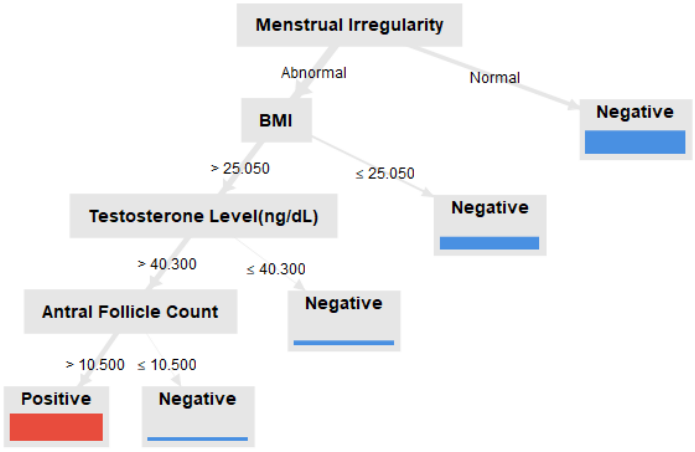
Setelah hasil dari *confusion matrix* diperoleh, langkah berikutnya adalah menghitung *performance vector*, yang merupakan sekumpulan metrik penilaian yang digunakan untuk mengevaluasi kinerja model klasifikasi secara menyeluruh. Tabel 8., memperlihatkan hasil dari perhitungan *performance vector.*

Tabel 7. *Performance Vector*

Performance Vector		
Accuracy	***Precision***	***Recall***
83%	94%	78%

7. Penggunaan *RapidMiner*

Proses pengolahan data dilakukan dengan menggunakan aplikasi *RapidMiner* untuk mengevaluasi apakah hasil yang dihitung secara manual sesuai dengan keluaran yang diperoleh dari sistem. Gambar 3., adalah hasil penggunaan *RapidMiner* yang menghasilkan diagram keputusan sebagai berikut.



Gambar 3. *Decision Tree RapidMiner.*

4. KESIMPULAN

Berdasarkan hasil pengamatan yang telah dilakukan terhadap data, dapat disimpulkan bahwa atribut *Menstrual Irregularity* merupakan faktor paling dominan (*root node*) dalam penentuan *diagnosis PCOS*. Atribut ini diikuti oleh *BMI*, *Testosterone Level* (*ng/dL*), dan *Antral Follicle Count* sebagai penentu selanjutnya dalam struktur pohon keputusan. Model yang dikembangkan mampu mencapai akurasi sebesar 83%, dengan *precision* mencapai 94% dan *recall* sebesar 78%. Hasil ini mengindikasikan bahwa model tersebut memiliki kemampuan yang sangat baik dalam mengenali kasus positif *PCOS* dengan tingkat kesalahan yang minim. Perbandingan antara hasil perhitungan manual menggunakan *Excel* dan proses otomatis melalui *RapidMiner* menghasilkan struktur pohon yang sama, sehingga menegaskan kredibilitas dan konsistensi metode yang

digunakan. Berbeda dari beberapa studi sebelumnya yang hanya fokus pada akurasi, penelitian ini juga menyajikan interpretasi yang jelas mengenai faktor-faktor utama penyebab *PCOS* melalui pohon keputusan. Hal ini dapat memberikan nilai tambah dalam meningkatkan pemahaman masyarakat dan tenaga medis, serta mendukung proses deteksi awal terhadap risiko *PCOS*.

DAFTAR PUSTAKA

[1] A. E. SusiloM. Di Lorenzo *et al.*, "Pathophysiology and Nutritional Approaches in Polycystic Ovary Syndrome (PCOS): A Comprehensive Review," *Curr. Nutr. Rep.*, vol. 12, no. 3, pp. 527–544, 2023, doi: 10.1007/s13668-023-00479-8.
[2] J. P. Christ and M. I. Cedars, "Current Guidelines for Diagnosing PCOS," Diagnostics, vol. 13, no. 6, 2023, doi: 10.3390/diagnostics13061113.
[3] W. H. O. WHO, "Polycystic ovary syndrome," Feb 7th. Accessed: Apr. 18, 2025. [Online]. Available: https://www-who-int.translate.goog/news-room/fact-sheets/detail/polycystic-ovary-syndrome?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc
[4] B. C. Sydora, M. S. Wilke, M. McPherson, S. Chambers, M. Ghosh, and D. F. Vine, "Challenges in diagnosis and health care in polycystic ovary syndrome in Canada: a patient view to improve health care," BMC Womens. Health, vol. 23, no. 1, pp. 1–18, 2023, doi: 10.1186/s12905-023-02732-2.
[5] W. Q. Salsabila, K. Adyani, and F. Realita, "Literatur Review: Faktor Resiko Sindrom Ovarium Polikistik pada Remaja," J. Heal., vol. 11, no. 02, pp. 164–174, 2024, doi: 10.30590/joh.v11n2.832.
[6] D. Fitriani, Y. Wahyuni, and R. Nuzrina, "Hubungan Status Gizi, Riwayat Siklus Menstruasi, dan Tingkat Depresi Terhadap Kejadian Polycystic Ovary Syndrome pada Wanita Usia Subur di RSAB Harapan Kita," Darussalam Nutr. J., vol. 7, no. 2, pp. 139–148, 2023, doi: 10.21111/dnj.v7i2.10721.
[7] V. Risdiyaningsih, E. Y. Kurniawati, and D. Darmawati, "Faktor Risiko Terjadinya Sindrom Ovarium Polikistik (Sopk)," J. Ilmu Kebidanan, vol. 9, no. 2, pp. 107–111, 2023, doi: 10.48092/jik.v9i2.209.
[8] N. Nur Melati Tanjung and A. Fauzi, "Hubungan Antara Kejadian Polycystic Ovarium Syndrome Dengan Akne Pada Wajah di NU Beauty Medical Aestetics Jonggol," J. Ilm. Keperawatan, vol. 9, no. 3, pp. 74–82, 2023.
[9] F. Novianti and N. Ulinnuha, "Seleksi Fitur Algoritma Genetika Dalam Klasifikasi Data Rekam Medis PCOS Menggunakan SVM," vol. 9, no. 1, pp. 9–19, 2024.
[10] J. Leadership, S. Di, and S. M. A. Muhammadiyah, "Jurnal Pendidikan Inklusif," vol. 8, no. 2, pp. 46–54, 2024.
[11] M. K. Afkar and M. Wali, "Aplikasi Prediksi Produksi Cabai dengan Algoritma C . 45 untuk Dinas Pertanian Provinsi Aceh Berbasis Web Abstrak," J. Ilmu Komput. dan Teknol. Inf., vol. 1, no. 1, 2024.
[12] N. T. Pitaloka and K. Kusnawi, "Pcos Disease Classification Using Feature Selection Rfecv and Eda With Knn Algorithm Method," J. Tek. Inform., vol. 4, no. 4, pp. 693–701, 2023, doi: 10.52436/1.jutif.2023.4.4.831.
[13] Suyanto, "Machine Learning tingkat dasar dan lanjut," in UBSI 07, 2018.
[14] I. Carolina, B. Lubis, A. Supriyatna, and R. Komarudin, "Application K-Nearest Neighbor Method with Particle Swarm Optimization for Classification of Heart Disease," no. April 2024, pp. 181–186, 2024, doi: 10.5220/0012446200003848.
[15] A. Praditya, M. R. Zein, and M. Nadriani, "Klasifikasi Potensi Turnover Karyawan Berdasarkan Data Kinerja Menggunakan Algoritma C4.5 Studi," vol. 9, no. 2, pp. 2723–2730, 2025