Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 1 (2025) pp: 144-149
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analisis Data Sintetis *Polycystic Ovary Syndrome* Menggunakan Algoritma Naive Bayes dan K-NN

Sigit Andriyanto[1], Sita Muharni[2], Sulistiyanto[3]
[1]Rekam Medis dan Informasi Kesehatan, Fakultas Kesehatan, Universitas Aisyah Pringsewu
[2]Sistem Informasi, Fakultas Teknik Bisnis dan Sains, Universitas Dharma Wacana
[3]Manajemen Informatika, Politeknik Negeri Sriwijaya
sigitandriyanto@aisyahuniversity.ac.id[1]*, sitamuharni@dharmawacana.ac.id[2], sulistiyanto@polsri.ac.id[3]

Abstrak

Kesehatan reproduksi wanita adalah aspek penting yang memiliki dampak besar pada kualitas hidup, baik secara fisik, emosional dan sosial. Kondisi yang sering mengganggu keseimbangan kesehatan reproduksi wanita adalah sindrom ovarium polikistik atau disebut PCOS (Polycystic Ovary Syndrome). PCOS adalah gangguan hormonal yang cukup kompleks dan sering ditemukan pada wanita. Diperkirakan sekitar 5 hingga 10% wanita dari usia remaja menghasilkan kondisi ini, menjadikannya salah satu gangguan endokrin yang paling umum pada wanita. Dalam konteks kesehatan, ini dapat berarti mengklasifikasikan apakah seseorang mungkin memiliki penyakit atau tidak sesuai dengan data gejala atau hasil tes laboratorium, Data sintetis yang mewakili kasus PCOS. Studi ini menggunakan dua algoritma klasifikasi yang akan digunakan untuk menganalisis data sintetis PCOS. Total data dari 3000 reocrd dapat ditentukan bahwa keakuratan menunjukkan bahwa keakuratan kisaran data dari data sindrom ovarium polifoni (PCOS) 93.83% untuk algoritma naive bayes, sedangkan untuk data yang diprediksi diagnosa PCOS (Polycystic Ovary Syndrome) menghasilkan 99.33% untuk algoritma K-NN. Oleh karena itu, dapat disimpulkan bahwa algoritma K-NN dapat menghitung dengan baik dengan menggunakan data sintetis wanita dengan PCOS (Polycystic Ovary Syndrome).

Kata kunci: Klasifikasi, PCOS, Naive Bayes, K-NN, Data Mining.

1. Latar Belakang

Kesehatan reproduksi perempuan merupakan aspek penting yang berdampak besar pada kualitas hidup, baik secara fisik, emosional, maupun sosial. Salah satu kondisi yang kerap mengganggu keseimbangan kesehatan reproduksi wanita adalah Sindrom Ovarium Polikistik atau yang lebih dikenal dengan PCOS (Polycystic Ovary Syndrome)[1]. PCOS adalah gangguan hormonal yang cukup kompleks dan sering dijumpai pada wanita usia subur. Diperkirakan, sekitar 5–10% wanita usia produktif mengalami kondisi ini, menjadikannya salah satu gangguan endokrin paling umum pada perempuan[2]. PCOS bukan hanya tentang menstruasi yang tidak teratur atau kesulitan untuk hamil[3]. Lebih dari itu, PCOS juga berkaitan erat dengan peningkatan kadar hormon androgen (hormon laki-laki), resistensi insulin, obesitas, jerawat parah, pertumbuhan rambut berlebih, hingga risiko jangka panjang seperti diabetes tipe 2, gangguan jantung, dan bahkan depresi[4]. Sayangnya, karena gejalanya sangat beragam dan sering tumpang tindih dengan kondisi medis lainnya, PCOS sering kali tidak terdeteksi dengan cepat. Banyak wanita yang baru menyadari bahwa mereka mengidap PCOS setelah bertahun-tahun mengalami gejala yang membingungkan[1].

Klasifikasi adalah proses pengelompokan data berdasarkan karakteristik tertentu. Dalam konteks medis, ini bisa berarti mengklasifikasikan apakah seseorang berisiko mengidap suatu penyakit atau tidak berdasarkan data gejala atau hasil pemeriksaan laboratorium[5]. Metode ini dapat membantu tenaga medis dalam mengambil keputusan, mempercepat proses diagnosis, dan bahkan meningkatkan akurasi dalam identifikasi penyakit. Namun, implementasinya dalam dunia nyata tidak selalu mudah[6]. Tantangan utama yang sering dihadapi adalah keterbatasan data, terutama data medis yang sangat sensitif, sulit diakses, dan harus dijaga kerahasiaannya. Untuk mengatasi kendala tersebut, penggunaan data sintetis menjadi solusi yang semakin populer[7]. Data sintetis adalah data yang dihasilkan secara artifisial menggunakan teknik tertentu, namun tetap merepresentasikan pola dan karakteristik yang mirip dengan data nyata[8]. Dalam penelitian dan pengembangan sistem berbasis machine learning, data sintetis bisa menjadi "arena latihan" yang sangat berguna terutama ketika data asli belum bisa

digunakan. Dengan menggunakan data sintetis yang mewakili kasus PCOS, kita bisa menguji efektivitas berbagai algoritma klasifikasi tanpa harus menunggu akses data riil dari rumah sakit atau laboratorium.

Dalam penelitian ini, dua algoritma klasifikasi yang akan digunakan untuk menganalisis data sintetis PCOS adalah Naive Bayes dan K-Nearest Neighbor (K-NN)[9]. Keduanya merupakan algoritma yang sudah cukup lama dikenal dan terbukti efektif dalam banyak kasus klasifikasi, termasuk di bidang medis[9]. Dengan membandingkan performa kedua algoritma ini pada data sintetis PCOS, penelitian ini bertujuan untuk mengevaluasi efektivitas dan akurasi masing-masing dalam mengklasifikasikan kondisi tersebut[10]. Proses ini bukan hanya penting dari sisi teknis, tetapi juga bisa menjadi pijakan awal bagi pengembangan sistem pendukung diagnosis yang lebih cerdas, efisien, dan terjangkau[11].

1. Metode Penelitian

Dalam metode pengumpulan data ini mempunyai peranan yang penting untuk mendapatkan suatu informasi dari penelitian yang dilakukan. Data yang relevan dengan pokok pembahasan adalah indikator keberhasilan penelitian. Pengumpulan data harus dilakukan dengan cara yang sangat tepat. Dalam metode pengumpulan data ini, penulis menggunakan beberapa metode yaitu :

2.1. Data *Understanding*

Tahap data *selection* atau pemilihan data merupakan tahap pemilihan atribut dari data yang akan dianalisis, karena tidak semua data yang terdapat dalam data mentah akan digunakan. Sehingga didapat beberapa atribut yang akan digunakan. Hasil tabel data selection dapat dilihat pada Tabel 1

Tabel 1 Data *Selection*

Atribut	No
	Jenis Kelamin
	Asuransi
	Usia
	Indeks Masa Tubuh
	Haid Tidak Teratur
	Tingkat Testoterone(ng/dL)
	Jumlah Antral Follicele

Data yang digunakan adalah data sekunder yaitu hasil yang didapat dari rekam medis puskesmas di dalam data tersebut dapat diketahui klasifikasi yang terbentuk yang terdiri dari 8 atribut prediksi dan 5 atribut target. Atribut-atribut yang menjadi parameter Tabel dibawah menerangkan atribut yang akan digunakan dan tidak digunakan dalam penelitian ini. Indikator "√" adalah menandakan bahwa atribut tersebut dieliminasi atau tidak digunakan pada tahap penentuan kriteria. Pengeliminasian beberapa atribut tersebut karena tidak mempengaruhi hasil dari proses penilaian pada Tabel 2.

Tabel 2 Atribut Data

Atribut	Indikator	Detail Penggunaan
No	-	Tidak digunakan
Jenis Kelamin	-	Tidak digunakan
Asuransi	-	Tidak digunakan
Usia	√	Digunakan
Indeks Masa Tubuh	√	Digunakan
Haid Tidak Teratur	√	Digunakan
Tingkat Testoterone(ng/dL)	√	Digunakan
Jumlah Antral Follicele	√	Digunakan

2.2. *Pemodelan*

Pada tahap ini sumber data yang dihasilkan berupa dataset dengan jumlah 3000 record data yang berisi data PCOS (*Polycystic Ovary Syndrome*). Kemudian data tersebut akan dijadikan landasan awal proses mining data sehingga didapatkan suatu output signifikan dan analisa hasil yang menyatakan algoritma yang lebih dominan dalam prediksi.

Tabel 3 Contoh data set dari 3000 record berisi data Sintetis untuk PCOS (*Polycystic Ovary Syndrome*)

Usia	Indeks Masa Tubuh	Haid Tidak Lancar	Tingkat Testoterone(ng/dL)	Jumlah Antral Follicele
29	21	0	46	9
20	21	0	59	6
23	23	0	69	10
19	33	1	78	37

DOI: https://doi.org/10.31004/riggs.v4i1.387

19	26	0	49	5
…	…	…	…	…

2.3. Implementasi Pengujian Algoritma

Pada tahap Implementasi pengujian klasterisasi pada penelitian ini menggunakan software RapidMiner. RapidMiner merupakan pemrograman lunak yang bekerja dalam pengolahan data. RapidMiner merupakan bahasa pemrograman yang mempunyai cakupan kemampuan yang luas dengan menggunakan prinsip dan algoritma data mining. Disamping itu RapidMiner dapat mengekstrakan pola-pola.

2.4. Evaluation

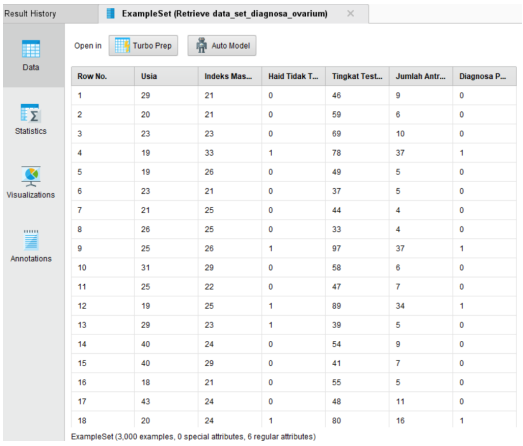
Pada tahap ini akan dilakukan analisa atau pengukuran ketepatan terhadap pemodelan yang telah dilakukan. Evaluasi ditujukan untuk mengetahui pemodelan yang dilakukan apakah sudah tepat dan sesuai diterapkan pada kasus penelitian ini serta sudah sesuai rencana awal penelitian. Selanjutnya dari hasil evaluasi tersebut adalah menentukan langkah berikutnya apakah bisa dilanjutkan atau diulang dari awal karena tidak sesuai dengan rencana awal.penelitian.

3. Hasil dan Diskusi

Pada penelitian ini metode yang digunakan adalah K-NN Nearst Neighbors and Naive Bayes Algorithm Clasification. Data yang digunakan ialah data rekam medis puskesmas tentang kumpulan data sintetis wanita dengan PCOS (*Polycystic Ovary Syndrome*) cara ekstensi rapidminner.

3.1. Data *Understanding*

Pengumpulan data awal mencakup pengumpulan data yang diperlukan untuk mendukung pemahaman data. Sumber utama data yang digunakan dalam penelitian ini adalah data sintetis wanita dengan PCOS (*Polycystic Ovary Syndrome*). Selain itu, data dibuat untuk membersihkan sebelum label untuk mengatur data file CSV terbaru yang dilindungi dari duplikasi data dan tanda -tanda yang tidak perlu. Setelah itu, langkah terakhir dalam mengumpulkan data, yaitu pelabelan, adalah kalsifikasi otomatis. Berikut hasil dari proses pengumpulan data ditunjukan pada gambar 1.

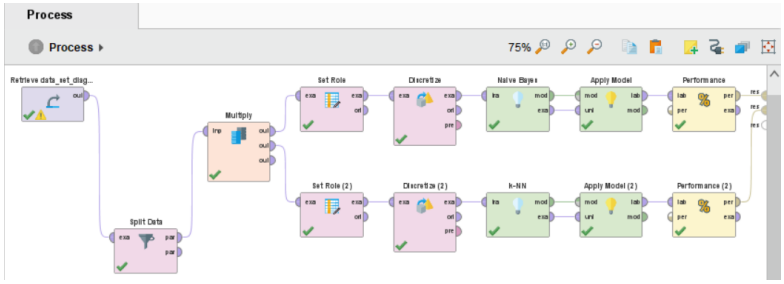


Gambar 1. Dataset PCOS (*Polycystic Ovary Syndrome*).

3.2. Pemodelan dan Implementasi Pengujian Algoritma

Pada tahap ini, peneliti akan mengevaluasi perhitungan angka untuk keperluan revisi nanti. Setelah berhasil dalam adegan, set data memiliki rekor 3000 yang cocok untuk digunakan. Untuk langkah pemodelan, peneliti akan mengukur kinerja klasifikasi dengan memodelkan dengan menggunakan bagian dari 2, yaitu 0,8 dan 0,2 menggunakan split data. Berikut ini adalah proses moddeling yang dibuat dalam perangkat lunak RapidMinner yang diilustrasikan pada Gambar 2.

DOI: https://doi.org/10.31004/riggs.v4i1.387



Gambar 2.Model Proses Modelling dan Implementasi Pengujian Algoritma

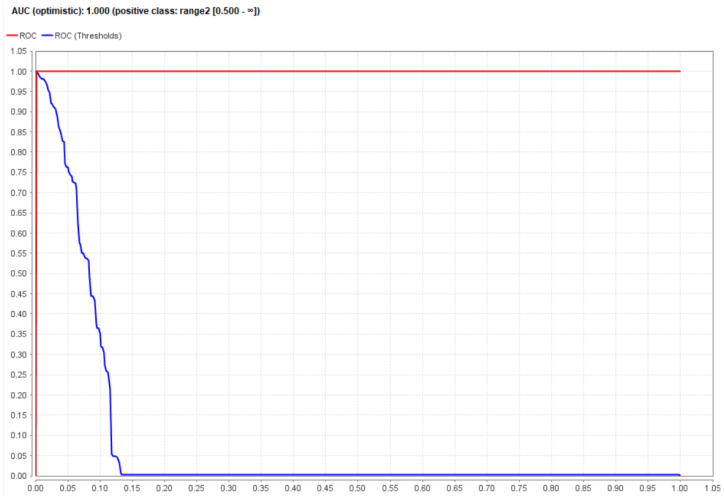
a. Hasil akurasi *confusin matrix* dan grafik *Area Under Curve* Naive Bayes
Berikut ini adalah Tabel hasil akurasi *confussion matrix* dan grafik *area under curve* Naive Bayes dalam analisis data sintetis wanita dengan PCOS (*Polycystic Ovary Syndrome*).

Table View Plot View

accuracy: 93.83%

	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred. range1 [-∞ - 0.500]	422	0	100.00%
pred. range2 [0.500 - ∞]	37	141	79.21%
class recall	91.94%	100.00%	

Gambar 3. Hasil Akurasi *Confusion Matrix*



Gambar 4. Grafik *Area Under Curve* (AUC)

b. Hasil akurasi *confusin matrix* dan grafik *Area Under Curve* K-NN

Berikut ini adalah Tabel hasil akurasi *confussion matrix* dan grafik *area under curve* K-NN dalam analisis data sintetis wanita dengan PCOS (*Polycystic Ovary Syndrome*).

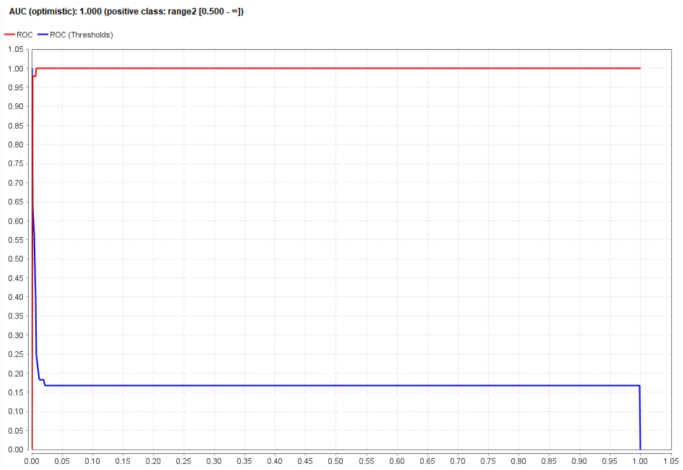
Table View Plot View

accuracy: 99.33%

	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred. range1 [-∞ - 0.500]	458	3	99.35%
pred. range2 [0.500 - ∞]	1	138	99.28%
class recall	99.78%	97.87%	

Gambar 5. Hasil Akurasi *Confusion Matrix*

DOI: https://doi.org/10.31004/riggs.v4i1.387



Gambar 6. Grafik *Area Under Curve* (AUC)

3.3. Evaluation

Pengujian menggunakan *confusion matrix* untuk melihat hasil pengujian data yang diperoleh dari tahapan modelling dengan mengunakan algoritma Naive Bayes dan K-NN. Total dataset 3000 record dapat dibuat kesimpulan bahwa precision menunjukkan tingkat ketepatan data yang diprediksi diagnosa PCOS (*Polycystic Ovary Syndrome*) range 1 menghasilkan presntase ketepatannya 91.94%, sedangkan untuk data range 2 sebesar 100% untuk algoritma naive bayes dan tingkat akurasi 93.83%, sedangkan untuk data yang diprediksi diagnosa PCOS (*Polycystic Ovary Syndrome*) range 1 menghasilkan presntase ketepatannya 99.78%, sedangkan untuk data range 2 sebesar 97.67% untuk algoritma K-NN dan tingkat akurasi 99.33%. Sehingga dapat disimpulkan bahwa algoritma K-NN dapat mengkalsifikasikan dengan baik menggunakan data sintetis wanita dengan PCOS (*Polycystic Ovary Syndrome*).

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan menegenai analisis data sintetis wanita dengan sindrom PCOS (*Polycystic Ovary Syndrome*) yaitu memprediksi hasil diagnosa PCOS (*Polycystic Ovary Syndrome*) menggunakan algoritma Niave Bayes dan K-NN serta menggunakan *software* rapid minner dengan parameter usia, indeks masa tubuh, haid tidak teratur, tungkat *testoterone* (ng/dL), dan jumlah antral *Follicele.* Menghasilkan suatu prediksi yang memiliki gejala sindrom PCOS (*Polycystic Ovary Syndrome*) dengan melihat hasil prediksi diganosa sindrom PCOS (*Polycystic Ovary Syndrome*) menggunakan algoritma K-NN merupakan algoritma yang akurat dalam memprediksi berbanding dengan algoritma Naive Bayes. Berdasarkan hasil pengujian menggunakan rapid minner maka didapatkan hasil terbaik dari algoritma K-NN yakni hasil akurasi 99.33%

Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada Universitas Aisyah Pringsewu dan dosen kolaborasi perguruan tinggi yang telah memberi bimbingan dan membantu publikasi terhadap penelitian ini.

Referensi

[1] N. T. Pitaloka and K. Kusnawi, “Pcos Disease Classification Using Feature Selection Rfecv And Eda With KNN Algorithm Method,” *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 4, pp. 693–701, 2023.
[2] A. E. Joham *et al.*, “Polycystic ovary syndrome,” *Lancet Diabetes Endocrinol*, vol. 10, no. 9, pp. 668–680, 2022.
[3] E. Stener-Victorin *et al.*, “Polycystic ovary syndrome,” *Nat Rev Dis Primers*, vol. 10, no. 1, p. 27, 2024.
[4] N. A. Stańczak, E. Grywalska, and E. Dudzińska, “The latest reports and treatment methods on polycystic ovary syndrome,” *Ann Med*, vol. 56, no. 1, p. 2357737, 2024.
[5] Y. Che, J. Yu, Y.-S. Li, Y.-C. Zhu, and T. Tao, “Polycystic ovary syndrome: challenges and possible solutions,” *J Clin Med*, vol. 12, no. 4, p. 1500, 2023.
[6] S. Sulistiyanto, E. Nadeak, N. Rahmi, and M. Malahayati, “Metode Data Mining dalam Kasus Seleksi Beasiswa: Literature Review,” *Jurnal Penelitian Inovatif*, vol. 4, no. 3, pp. 1091–1100, 2024.

DOI: https://doi.org/10.31004/riggs.v4i1.387

[7] S. Vinothini, S. Vaishnavi, and N. Mythili, "Polycystic Ovary Syndrome (PCOS) Disease Prediction Using Machine Learning," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, IEEE, 2024, pp. 1–9.

[8] S. Muharni and S. Andriyanto, "Penentuan Pola Penjualan Menggunakan Algoritma Apriori," *Digital Transformation Technology*, vol. 4, no. 1, pp. 60–71, 2024.

[9] N. Nosiel, S. Andriyanto, and M. S. Hasibuan, "Application of Nave Bayes Algorithm for SMS Spam Classification Using Orange," *International Journal of Advanced Science and Computer Applications*, vol. 1, no. 1, pp. 16–24, 2022.

[10] M. S. Hasibuan, R. Z. A. Aziz, and A. Sigit, "Utilizing Clustering Algorithms to Provide Vark Learning Style Recommendations," in *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, IEEE, 2023, pp. 361–365.

[11] S. Ahmed *et al.*, "A review on the detection techniques of polycystic ovary syndrome using machine learning," *IEEE Access*, vol. 11, pp. 86522–86543, 2023.

DOI: https://doi.org/10.31004/riggs.v4i1.387