JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Analisis Perbandingan Optimasi Seleksi Fitur Logistic Regression dan SVM untuk Prediksi PCOS

Annisa Ashari[1], Lumi Krismona[2], Nurhayati[3]
[1,2,3] Ilmu Komputer, Universitas Potensi Utama
Email: [1]annisaashari19@gmail.com, [2]lumiikrismona@gmail.com, [3]n.hayati2390@gmail.com
Email Penulis Korespondensi: annisaashari19@gmail.com

Abstrak
Polycystic Ovary Syndrome (PCOS) merupakan gangguan endokrin pada wanita usia reproduktif yang ditandai ketidakteraturan menstruasi, hiperandrogenisme dan perubahan morfologi ovarium, serta berpotensi menimbulkan infertilitas dan komplikasi metabolik (WHO, 2025). Penelitian ini bertujuan menganalisis secara komparatif kinerja algoritma *Logistic Regression* dan *Support Vector Machine* (SVM) yang dioptimasi dengan *feature selection* untuk prediksi PCOS berbasis data klinis. Dataset berisi 1.000 data pasien dengan lima atribut klinis, yaitu umur, indeks massa tubuh (BMI), ketidakteraturan menstruasi, kadar testosteron dan jumlah folikel antral, serta label biner diagnosis PCOS. Data dibagi menggunakan *stratified train-test split* 80:20 dan seluruh fitur numerik dinormalisasi. Optimasi dilakukan dengan mengintegrasikan *Recursive Feature Elimination* (RFE) dan *Grid Search* pada *Logistic Regression* untuk menentukan kombinasi jumlah fitur dan parameter regulasi terbaik, sementara pada SVM dilakukan penalaan parameter C, jenis kernel dan gamma menggunakan *GridSearchCV* dengan *5-fold stratified cross-validation* dan *F1-score* sebagai metrik acuan, mengikuti praktik optimasi model yang banyak digunakan pada penelitian PCOS berbasis *machine learning*. Hasil eksperimen menunjukkan bahwa *Logistic Regression* terbaik mencapai akurasi 0,915, *F1-score* kelas PCOS 0,80 dan AUC 0,978, sedangkan SVM memberikan kinerja lebih tinggi dengan akurasi 0,97, *F1-score* kelas PCOS 0,92 dan AUC 0,998. Secara keseluruhan, hasil ini mengindikasikan bahwa SVM dengan fitur terpilih lebih efektif dibanding *Logistic Regression*, selaras dengan beberapa studi yang melaporkan keunggulan model SVM dalam deteksi PCOS.
Kata Kunci: *Polycystic Ovary Syndrome*, *Logistic Regression*, *Support Vector Machine*, *Feature Selection*, *Machine Learning*.

Abstract
Polycystic Ovary Syndrome (PCOS) is an endocrine disorder in reproductive-age women characterized by menstrual irregularities, hyperandrogenism, and ovarian morphological changes, and is associated with infertility and metabolic complications (WHO, 2025). This study aims to comparatively analyze the performance of Logistic Regression and Support Vector Machine (SVM) algorithms optimized with feature selection for PCOS prediction based on clinical data. The dataset consists of 1,000 patient records with five clinical attributes, namely age, body mass index (BMI), menstrual irregularity, testosterone level, and antral follicle count, as well as a binary label indicating PCOS diagnosis. The data are split using an 80:20 stratified train–test scheme, and all numerical features are normalized. Model optimization is carried out by integrating Recursive Feature Elimination (RFE) and Grid Search on Logistic Regression to determine the best combination of number of features and regularization parameters, while SVM hyperparameters (C, kernel type, and gamma) are tuned using GridSearchCV with 5-fold stratified cross-validation and F1-score as the objective metric, following common optimization practices in PCOS machine learning studies. Experimental results show that the best Logistic Regression model achieves an accuracy of 0.915, PCOS-class F1-score of 0.80, and AUC of 0.978, whereas the best SVM model attains a higher accuracy of 0.97, PCOS-class F1-score of 0.92, and AUC of 0.998. Overall, these findings indicate that SVM with selected features is more effective than Logistic Regression on the given PCOS dataset, in line with several studies reporting the superiority of SVM models for PCOS detection.
Keywords*: Polycystic Ovary Syndrome, Logistic Regression, Support Vector Machine, Feature Selection, Machine Learning.*

1. PENDAHULUAN

Polycystic Ovary Syndrome (PCOS) merupakan gangguan endokrin pada wanita usia reproduksi yang ditandai gangguan ovulasi, hiperandrogenisme dan morfologi ovarium polikistik, serta berasosiasi dengan infertilitas dan peningkatan risiko sindrom metabolik, diabetes dan penyakit kardiovaskular. Meskipun prevalensinya tinggi, PCOS sering terlambat terdiagnosis karena spektrum gejala yang luas dan bergantung pada interpretasi klinisi, sehingga banyak kasus baru teridentifikasi setelah muncul masalah kesuburan atau komorbiditas lain [1]. Kondisi ini menimbulkan kebutuhan pendekatan skrining yang lebih objektif dan terstandardisasi berbasis data klinis terstruktur, yang dapat membantu tenaga kesehatan dalam deteksi dini PCOS [2] .

Perkembangan *machine learning* (ML) beberapa tahun terakhir memberikan alternatif untuk membangun model prediksi PCOS dengan memanfaatkan kombinasi fitur klinis, laboratorik dan gaya hidup [3]. Berbagai studi melaporkan bahwa algoritma klasifikasi seperti *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbor*, serta model *ensemble* mampu mencapai akurasi di atas 90% pada tugas klasifikasi PCOS ketika dipadukan dengan teknik seleksi fitur dan pra-pemrosesan yang tepat [4]. Namun, pemilihan kombinasi algoritma dan strategi optimasi yang paling seimbang antara performa, kompleksitas dan interpretabilitas pada skenario klinis praktis masih menjadi topik penelitian yang terbuka [5].

JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Sejumlah penelitian telah fokus pada deteksi PCOS berbasis ML [6] memanfaatkan *Exploratory Data Analysis* (EDA) dan SVM dengan berbagai kernel pada dataset PCOS Kaggle berisi lima fitur klinis dan mendapatkan akurasi terbaik 89,62% dengan kernel polinomial sehingga menunjukkan kemampuan SVM menangani distribusi data yang tidak seimbang dan pola non-linier sederhana [7] mengombinasikan EDA dan metode seleksi fitur RFECV dengan algoritma *K-Nearest Neighbor* dan berhasil mereduksi fitur menjadi tujuh atribut paling penting sekaligus meningkatkan akurasi klasifikasi hingga 93%, menegaskan dampak positif *feature selection* terhadap kualitas model PCOS. [8] mengembangkan pendekatan prediksi PCOS berbasis web dengan 13 algoritma ML dan *feature selection Mutual Information*; *Random Forest* dan *AdaBoost* mencapai akurasi 94% dan ditunjukkan bahwa pemilihan fitur dominan berperan penting dalam stabilitas dan generalisasi model.

Analisis komparatif beberapa algoritma ML (KNN, *Naive Bayes*, SVM, *Decision Tree* dan *Logistic Regression*) pada data survei gejala PCOS dan menemukan bahwa *Decision Tree* memberikan akurasi tertinggi 81%, sementara *Logistic Regression* dan SVM berperan sebagai *baseline* linier yang penting untuk dibandingkan dari sisi kesederhanaan dan interpretabilitas. [1] memanfaatkan 30.601 rekam medis elektronik dan membangun model prediksi PCOS menggunakan *Logistic Regression*, SVM, *Gradient Boosted Trees* dan *Random Forest*; model terbaik mencapai AUC hingga 0,85 dan menunjukkan bahwa kombinasi fitur klinis dan hormonal yang dipilih secara sistematis dapat meningkatkan kinerja model untuk deteksi dini PCOS di populasi berisiko menunjukkan bahwa integrasi beberapa teknik seleksi fitur (filter, *embedded* dan *wrapper*) pada dataset PCOS Kaggle memungkinkan *Random Forest* mencapai akurasi hampir 99%, menguatkan pentingnya tahap *feature selection* sebelum pelatihan *classifier*.

Walaupun banyak studi telah mengkaji prediksi PCOS dengan beragam algoritma dan strategi seleksi fitur, kajian yang secara spesifik menyoroti perbandingan mendalam antara *Logistic Regression* dan SVM setelah proses *feature selection* yang terintegrasi pada dataset PCOS dengan fitur klinis yang relatif ringkas masih terbatas [9]. Sebagian besar penelitian berfokus pada model ensemble atau kombinasi banyak *classifier* tanpa menguraikan secara rinci bagaimana pengaruh pemilihan subset fitur terhadap kinerja dan karakteristik dua model dasar tersebut, terutama dalam konteks *trade-off* antara interpretabilitas dan kemampuan memodelkan pola non-linier [10].

Berdasarkan celah tersebut, penelitian ini bertujuan melakukan analisis komparatif kinerja *Logistic Regression* dan SVM yang dioptimasi dengan *feature selection* untuk prediksi PCOS berbasis lima fitur klinis utama, yaitu umur, indeks massa tubuh, ketidakteraturan menstruasi, kadar testosteron dan jumlah folikel antral. Optimasi dilakukan melalui kombinasi *Recursive Feature Elimination* (RFE) dan pencarian *hyperparameter* menggunakan *GridSearchCV* dengan skema *stratified k-fold cross-validation*, sehingga diperoleh konfigurasi fitur dan parameter terbaik untuk masing-masing algoritma. Penelitian ini diharapkan dapat menghasilkan model prediksi PCOS yang akurat dan tetap interpretabel, serta memberikan bukti empiris mengenai pengaruh *feature selection* terhadap kinerja *Logistic Regression* dan SVM pada dataset PCOS klinis sehingga dapat menjadi dasar pengembangan sistem pendukung keputusan untuk skrining dini PCOS di layanan kesehatan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian pada studi ini disusun secara sistematis agar proses *data mining* untuk prediksi PCOS berjalan terstruktur dan dapat direplikasi. Secara ringkas, tahapan tersebut adalah sebagai berikut :

1. Pengumpulan dan Pemahaman Data
   Mengambil dataset PCOS berisi lima atribut klinis (umur, BMI, ketidakteraturan menstruasi, kadar testosteron, jumlah folikel antral) dan satu label PCOS_Diagnosis, kemudian melakukan analisis deskriptif awal serta pengecekan distribusi kelas.
2. Pra-Pemrosesan Data
   Membersihkan data (cek nilai hilang/duplikat), melakukan *stratified train-test split* 80:20, lalu menstandardisasi seluruh fitur numerik agar siap digunakan oleh algoritma *Logistic Regression* dan SVM.
3. Seleksi Fitur (*Feature selection*)
   Menerapkan *Recursive Feature Elimination* (RFE) di dalam *pipeline Logistic Regression* untuk menentukan subset fitur terbaik, sekaligus menguji beberapa jumlah fitur yang dipertahankan.
4. Pembangunan dan Optimasi Model
   Melatih dua model utama, yaitu *Logistic Regression* dan SVM (kernel linear dan RBF), dengan optimasi *hyperparameter* (C, *penalty* pada LR; C, kernel, gamma pada SVM) menggunakan *cross-validation* dan *Grid Search* berbasis *F1-score*.
5. Evaluasi dan Analisis Hasil
   Mengukur kinerja model terbaik menggunakan akurasi, *precision*, *recall*, *F1-score*, AUC dan *Confusion matrix* pada data uji, kemudian membandingkan performa *Logistic Regression* dan SVM untuk menilai pengaruh seleksi fitur terhadap kemampuan prediksi PCOS.

JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Gambar 1. Tahapan Penelitian

2.2 Dataset

Dataset yang digunakan pada penelitian ini berupa data yang disusun sebagai data klinis terstruktur untuk kasus PCOS. Dataset berisi 1.000 baris data pasien wanita usia reproduktif. Setiap *record* terdiri dari lima atribut prediktor dan satu atribut target, yaitu: *Age* (tahun), *Body Mass Index*/BMI (kg/m²), *Menstrual_Irregularity* (0 = siklus teratur, 1 = tidak teratur), *Testosterone_Level* (ng/dL), *Antral_Follicle_Count* (jumlah folikel antral), serta label biner PCOS_Diagnosis yang menyatakan pasien terdiagnosis PCOS (1) atau tidak (0). Struktur fitur ini diselaraskan dengan penelitian sebelumnya yang menggunakan kombinasi parameter antropometri, status siklus menstruasi, kadar hormon dan karakteristik ovarium sebagai indikator utama risiko PCOS.

Dataset ini diadaptasi dari skema data PCOS yang juga digunakan pada beberapa penelitian terdahulu yang memanfaatkan kombinasi parameter umur, BMI, status ketidakteraturan menstruasi, kadar hormon dan karakteristik folikel sebagai fitur utama untuk prediksi PCOS. Data disimpan dalam format CSV agar mudah diproses menggunakan bahasa pemrograman Python, sesuai praktik umum pada studi prediksi PCOS berbasis *machine learning*. Melihat beberapa baris pertama (*head*), statistik deskriptif (rata-rata, minimum, maksimum dan simpangan baku), serta distribusi label PCOS_Diagnosis untuk mengetahui proporsi kelas PCOS dan non-PCOS. Menginterpretasikan masing-masing fitur sebagai faktor risiko klinis: BMI dan kadar testosteron yang tinggi, ketidakteraturan menstruasi dan jumlah folikel antral yang meningkat telah dilaporkan berhubungan erat dengan diagnosis PCOS pada berbagai studi sebelumnya.

Tabel 1. Data Set

Age	BMI	Menstrual_Irregularity	Testosterone_Level(ng/dL)	Antral_Follicle_Count	PCOS_Diagnosis
24	34.7	1	25.2	20	0
37	26.4	0	57.1	25	0
32	23.6	0	92.7	28	0
28	28.8	0	63.1	26	0
25	22.1	1	59.8	8	0

2.3 Pra-Pemrosesan

Pada tahap ini, kolom data yaitu *Age*, BMI, *Menstrual_Irregularity*, *Testosterone Level* (ng/dL), *Antral_Follicle_Count* dan PCOS_Diagnosis dicek keberadaan nilai kosong (*missing values*), nilai duplikat dan nilai yang berada di luar rentang wajar secara klinis, misalnya usia di bawah 18 tahun atau kadar testosteron jauh di luar kisaran yang dilaporkan pada studi PCOS [11]. Jika ditemukan baris dengan nilai kosong yang sedikit dan tidak kritis, data dapat dihapus; namun bila nilai kosong muncul pada variabel penting dan dalam jumlah kecil, biasanya diisi menggunakan nilai rata-rata (*mean*) atau median agar distribusi data tidak terlalu berubah, mengikuti pendekatan yang banyak digunakan pada penelitian PCOS berbasis Kaggle dan EHR. Seluruh atribut klinis diset sebagai tipe numerik (*integer* atau *float*), sedangkan PCOS_Diagnosis dipastikan hanya berisi nilai 0 dan 1 sehingga dapat diperlakukan sebagai variabel target biner dalam algoritma klasifikasi. Penyeragaman ini penting untuk mencegah kesalahan komputasi pada saat pemodelan, karena algoritma seperti *Logistic Regression* dan SVM mengharuskan input dalam bentuk matriks numerik dan label kelas yang konsisten. Selanjutnya pembagian data menjadi set pelatihan dan pengujian (*train-test split*). Untuk menjaga representativitas kelas, digunakan skema *stratified split* dengan komposisi 80% data sebagai data latih dan 20% sebagai data uji, sehingga proporsi pasien PCOS dan non-PCOS di kedua subset tetap seimbang. Pendekatan ini sejalan dengan

beberapa penelitian PCOS terkini yang menekankan pentingnya *stratification* pada dataset medis biner agar model tidak bias terhadap kelas mayoritas dan performa yang dilaporkan pada data uji benar-benar mencerminkan kemampuan generalisasi model.

Standardisasi fitur (*feature scaling*). Semua fitur numerik, yaitu *Age*, BMI, *Testosterone_Level* (ng/dL) dan *Antral_Follicle_Count*, diubah skalanya menggunakan teknik standardisasi (misalnya *StandardScaler*) sehingga tiap fitur memiliki rataan mendekati nol dan simpangan baku sekitar satu. Standardisasi ini sangat krusial untuk model yang berbasis jarak atau margin seperti SVM dan untuk algoritma yang menggunakan regularisasi seperti *Logistic Regression*, karena perbedaan skala antar fitur dapat menyebabkan satu variabel mendominasi proses pembelajaran dan mengganggu proses optimasi. Dengan skala yang seragam, pemilihan *hyperparameter* C dan gamma pada SVM maupun C dan penalti pada *Logistic Regression* menjadi lebih stabil dan konsisten antar *fold* pada *cross-validation*. Hasil pra-pemrosesan ini divalidasi kembali melalui ringkasan statistik dan visual sederhana (misalnya distribusi fitur setelah *scaling* dan distribusi label pada *train/test*) untuk memastikan tidak ada distorsi besar pada pola data asli, sebagaimana dianjurkan pada studi-studi PCOS yang mengombinasikan pra-pemrosesan, seleksi fitur dan pemodelan ML. Dengan demikian, dataset yang telah melalui tahapan pra-pemrosesan dapat digunakan secara andal pada tahap berikutnya, yaitu seleksi fitur dengan *Recursive Feature Elimination* dan pelatihan model *Logistic Regression* serta SVM yang dioptimasi.

2.4 Seleksi Fitur (*Feature Selection*)

Tahap seleksi fitur (*feature selection*) pada penelitian ini bertujuan memilih subset atribut klinis yang paling relevan terhadap diagnosis PCOS sehingga model menjadi lebih sederhana, akurat dan stabil. Dengan jumlah fitur yang relatif sedikit, seleksi fitur difokuskan untuk mengurangi redundansi, memperkuat sinyal kelas dan meningkatkan interpretabilitas model *Logistic Regression* maupun SVM. Metode utama yang digunakan adalah *Recursive Feature Elimination* (RFE) berbasis *Logistic Regression* yang diintegrasikan ke dalam *pipeline* bersama proses standardisasi dan *Grid Search hyperparameter*. Pada tahap ini, model *Logistic Regression* pertama-tama dilatih menggunakan semua fitur (*Age*, BMI, *Menstrual_Irregularity*, *Testosterone_Level*, *Antral_Follicle_Count*), kemudian RFE secara rekursif menghapus fitur dengan kontribusi paling kecil berdasarkan bobot (koefisien) model hingga tersisa jumlah fitur tertentu. Jumlah fitur yang akan dipertahankan tidak ditentukan secara tunggal sejak awal, tetapi diperlakukan sebagai parameter yang diuji, misalnya 2, 3, 4 dan 5 fitur, sehingga kombinasi "jumlah fitur terbaik + parameter regulasi C + jenis penalti (L1/L2)" untuk *Logistic Regression* dapat dicari secara bersamaan menggunakan *GridSearchCV* dengan skema *k-fold cross-validation* dan *F1-score* sebagai metrik tujuan [12].

Hasil dari proses RFE ini adalah subset fitur optimal yang memberikan kinerja terbaik pada validasi silang, misalnya kombinasi fitur BMI, *Menstrual_Irregularity*, *Testosterone_Level* dan *Antral_Follicle_Count* yang paling kuat dalam membedakan kelas PCOS dan non-PCOS. Subset fitur tersebut kemudian digunakan tidak hanya untuk melatih *Logistic Regression* teroptimasi, tetapi juga sebagai masukan ke model SVM (kernel linear maupun RBF) agar kedua algoritma diuji pada ruang fitur yang sama dan sudah tersaring dari atribut yang kurang informatif. Pendekatan ini sejalan dengan beberapa penelitian PCOS terkini yang menunjukkan bahwa kombinasi seleksi fitur berbasis model (seperti RFE, RFECV atau *Mutual Information*) dengan algoritma klasifikasi mampu meningkatkan akurasi dan *F1-score* sekaligus mengurangi risiko *overfitting* pada dataset klinis.

2.5 Pembangunan dan Optimasi Model

Tahap pembangunan dan optimasi model berfokus pada penyusunan dua model klasifikasi utama, yaitu *Logistic Regression* dan *Support Vector Machine* (SVM), serta penalaan *hyperparameter* untuk memperoleh konfigurasi dengan kinerja terbaik pada prediksi PCOS. Kedua model dibangun di atas data yang telah melalui tahapan pra-pemrosesan dan seleksi fitur, sehingga input yang digunakan sudah bersih, terstandardisasi dan terdiri dari atribut yang paling relevan terhadap diagnosis PCOS.

Pembangunan model dilakukan menggunakan *pipeline scikit-learn* yang menggabungkan *StandardScaler*, modul seleksi fitur (RFE untuk *Logistic Regression*) dan *classifier* di dalam satu alur pemrosesan. Untuk *Logistic Regression*, model dirancang sebagai klasifikator linier dengan fungsi *sigmoid* yang memetakan kombinasi linier fitur klinis menjadi probabilitas pasien mengalami PCOS, sehingga koefisien model dapat ditafsirkan sebagai kontribusi relatif masing-masing fitur. Pada model ini, parameter regulasi C (kekuatan regularisasi) dan jenis penalti (L1/L2) menjadi kunci untuk mengontrol kompleksitas dan menghindari *overfitting*, terutama pada dataset medis dengan jumlah fitur terbatas.

Pada SVM digunakan konfigurasi dengan kernel linear dan RBF guna membandingkan kemampuan pemisahan linier dan non-linier pada ruang fitur hasil seleksi. *Hyperparameter* utama yang dioptimasi adalah C (*trade-off* antara margin dan kesalahan klasifikasi) serta gamma untuk kernel RBF yang mengatur seberapa jauh pengaruh sebuah titik data terhadap margin keputusan. Kedua algoritma (*Logistic Regression* dan SVM) dioptimasi menggunakan *GridSearchCV* dengan skema *5-fold stratified cross-validation*, di mana kombinasi nilai C, jenis penalti (untuk LR), jenis kernel dan gamma (untuk SVM), serta jumlah fitur terpilih dari RFE dievaluasi menggunakan *F1-score* kelas PCOS sebagai metrik objektif utama.

Melalui proses ini diperoleh model *Logistic Regression* terbaik dan model SVM terbaik yang masing-masing sudah terkalibrasi pada kombinasi hyperparameter dan subset fitur yang memberikan performa tertinggi pada data latih, sebelum

JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

kemudian diuji pada data uji untuk menilai kemampuan generalisasinya. Pendekatan pembangunan dan optimasi berbasis *pipeline* dan *Grid Search* ini sejalan dengan praktik pada studi PCOS terbaru yang memadukan seleksi fitur dan penalaan *hyperparameter* untuk meningkatkan akurasi, *F1-score* dan AUC model klasifikasi.

2.6 Evaluasi dan Analisis Hasil

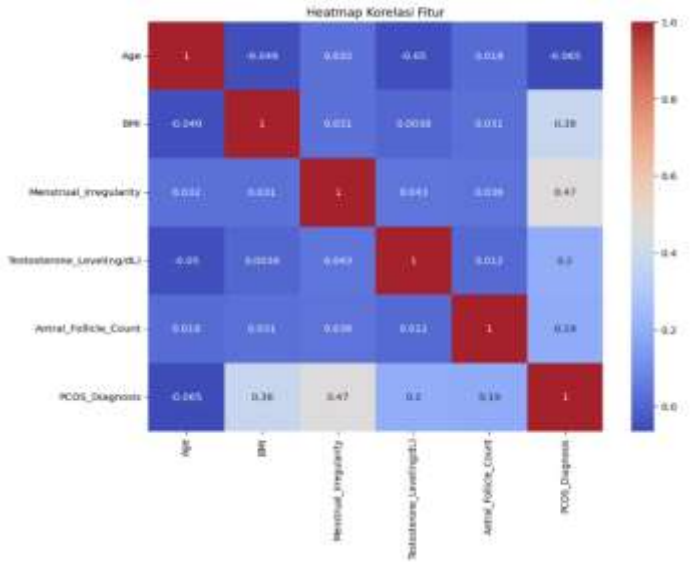
Secara kuantitatif, kinerja model diukur menggunakan beberapa metrik utama, yaitu *accuracy*, *precision*, *recall*, *F1-score* dan *area under the ROC curve* (AUC), yang umum digunakan dalam penelitian diagnosis PCOS berbasis *machine learning*. *Accuracy* digunakan untuk melihat proporsi keseluruhan prediksi yang benar, sementara *precision* dan *recall* fokus pada kualitas deteksi kelas positif (PCOS), yaitu seberapa banyak pasien yang diprediksi PCOS benar-benar PCOS dan seberapa banyak pasien PCOS yang berhasil terdeteksi. *F1-score* dihitung sebagai rata-rata harmonis *precision* dan *recall* untuk memberikan ukuran tunggal yang seimbang, sedangkan AUC digunakan untuk menilai kemampuan diskriminatif model di berbagai ambang keputusan, yang sangat penting dalam konteks skrining klinis.

Selain metrik agregat, disusun pula *confusion matrix* untuk kedua model (*Logistic Regression* terbaik dan SVM terbaik) yang memperlihatkan jumlah *true positive*, *true negative*, *false positive* dan *false negative* pada data uji. *Confusion matrix* ini dianalisis untuk memahami pola kesalahan model, terutama jumlah *false negative* (kasus PCOS yang tidak terdeteksi) yang secara klinis lebih kritis dibanding *false positive* dalam skenario skrining penyakit. Hasil menunjukkan bahwa SVM dengan fitur terpilih memberikan nilai akurasi, *F1-score* kelas PCOS dan AUC yang lebih tinggi dibanding *Logistic Regression*, serta menghasilkan jumlah *false positive* dan *false negative* yang lebih sedikit, sehingga dapat disimpulkan bahwa SVM lebih efektif untuk prediksi PCOS pada dataset ini. Secara kualitatif, analisis koefisien *Logistic Regression* dan fitur yang dipertahankan oleh RFE juga digunakan untuk menginterpretasikan fitur klinis yang paling berpengaruh dalam menentukan status PCOS dan temuan ini dibandingkan dengan literatur sebelumnya guna memastikan konsistensi dengan faktor risiko PCOS yang sudah dikenal.

3. HASIL DAN PEMBAHASAN

Pada bagian ini berisi hasil dan pembahasan dari topik penelitian, yang bisa di buat terlebih dahulu metodologi penelitian. Bagian ini juga merepresentasikan penjelasan yang berupa penjelasan, gambar, tabel dan lainnya.

3.1 Hasil Pra-Pemrosesan



Gambar 2. *Heatmap* Korelasi Fitur

Berdasarkan hasil analisis matriks korelasi dan visualisasi data, diperoleh gambaran hubungan antara variabel-variabel penelitian dengan diagnosis PCOS (*Polycystic Ovary Syndrome*) sebagai berikut.:

a. Variabel Usia (*Age*)
   Usia menunjukkan korelasi negatif yang sangat lemah terhadap PCOS_Diagnosis dengan nilai korelasi sebesar -0.065. Hasil ini mengindikasikan bahwa faktor usia tidak memiliki pengaruh yang berarti terhadap kemungkinan individu mengalami PCOS.
b. Indeks Massa Tubuh (BMI)
   Nilai korelasi antara BMI dan PCOS_Diagnosis adalah 0.38, yang menunjukkan adanya hubungan positif sedang. Dengan demikian, peningkatan indeks massa tubuh cenderung diikuti oleh peningkatan risiko

JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

terdiagnosis PCOS. Temuan ini sejalan dengan penelitian sebelumnya yang menyebutkan bahwa obesitas dapat memperburuk ketidakseimbangan hormonal dan mempertinggi insiden PCOS.

c. Ketidakteraturan Menstruasi (*Menstrual_Irregularity*)
Variabel ketidakteraturan menstruasi memiliki korelasi positif paling tinggi terhadap diagnosis PCOS, yaitu sebesar 0.47. Hal ini menunjukkan bahwa ketidakteraturan siklus menstruasi merupakan salah satu faktor prediktif utama PCOS. Kondisi ini sering dikaitkan dengan hiperandrogenisme yang berdampak pada gangguan ovulasi .

d. Kadar Testosteron (*Testosterone_Level*)
Korelasi antara kadar testosteron dan PCOS_Diagnosis sebesar 0.2, yang menunjukkan adanya hubungan positif namun relatif lemah. Meskipun demikian, kadar testosteron yang meningkat tetap dianggap sebagai salah satu indikator biologis penting dalam diagnosis PCOS.

e. Jumlah Folikel Antral (Antral_Follicle_Count)
Variabel jumlah folikel antral juga menunjukkan korelasi positif dengan PCOS_Diagnosis (0.19). Nilai ini menegaskan bahwa jumlah folikel antral yang tinggi merupakan karakteristik umum pada individu dengan PCOS.

Secara keseluruhan, hasil analisis menunjukkan bahwa variabel dengan korelasi tertinggi terhadap diagnosis PCOS adalah *Menstrual_Irregularity* (0.47) dan BMI (0.38). Hal ini menandakan dassiklus menstruasi yang tidak teratur dan peningkatan indeks massa tubuh merupakan dua faktor dominan yang berkaitan erat dengan PCOS dalam dataset penelitian ini

3.2 Evaluasi Hasil

```
--- Logistic Regression (best) ---
Accuracy : 0.915
Precision: 0.7555555555555555
Recall   : 0.85
F1-score : 0.8
AUC      : 0.9775

Classification report:
              precision    recall  f1-score   support

           0       0.96      0.93      0.95       160
           1       0.76      0.85      0.80        40

    accuracy                           0.92       200
   macro avg       0.86      0.89      0.87       200
weighted avg       0.92      0.92      0.92       200
```

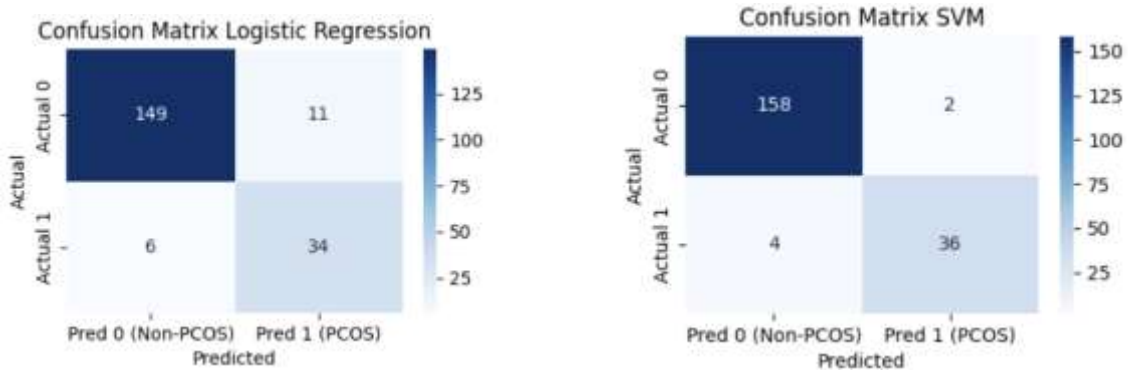
```
--- SVM (best) ---
Accuracy : 0.97
Precision: 0.9473684210526315
Recall   : 0.9
F1-score : 0.9230769230769231
AUC      : 0.99765625

Classification report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       160
           1       0.95      0.90      0.92        40

    accuracy                           0.97       200
   macro avg       0.96      0.94      0.95       200
weighted avg       0.97      0.97      0.97       200
```

Gambar 3. *Classification Report*

Berdasarkan hasil pengujian pada data uji, model *Logistic Regression* terbaik menghasilkan akurasi sebesar 0,915 dengan nilai *precision* kelas PCOS 0,76, *recall* 0,85, *F1-score* 0,80 dan AUC 0,9775. Nilai *precision* yang relatif lebih rendah dibanding *recall* menunjukkan bahwa masih terdapat sejumlah kasus non-PCOS yang salah diklasifikasikan sebagai PCOS (*false positive*), namun kemampuan model dalam menangkap kasus PCOS (*true positive*) sudah cukup baik, tercermin dari *recall* 0,85. Secara keseluruhan, *F1-score* dan AUC yang tinggi mengindikasikan bahwa *Logistic Regression* mampu memberikan pemisahan kelas yang cukup baik pada dataset ini dan dapat dijadikan baseline model yang interpretabel untuk menjelaskan pengaruh masing-masing fitur klinis terhadap risiko PCOS.

Sementara itu, model SVM terbaik menunjukkan kinerja yang lebih tinggi dengan akurasi 0,97, *precision* kelas PCOS 0,95, *recall* 0,90, *F1-score* 0,92 dan AUC 0,9977. Kombinasi *precision* dan *recall* yang sama-sama tinggi menggambarkan bahwa SVM tidak hanya mampu mengurangi jumlah *false positive*, tetapi juga menekan *false negative*, sehingga lebih andal untuk skenario skrining di mana kasus PCOS yang terlewat harus diminimalkan. Peningkatan *F1-score* dari 0,80 (*Logistic Regression*) menjadi 0,92 (SVM) serta AUC yang mendekati 1,0 menunjukkan bahwa SVM dengan konfigurasi dan fitur yang telah dioptimasi memberikan kemampuan diskriminatif yang sangat baik, sejalan dengan beberapa penelitian PCOS terkini yang juga melaporkan keunggulan SVM dan model non-linier lainnya dibanding model linier pada tugas klasifikasi PCOS.

JURNAL SISTEM INFORMASI TGD
Volume 5, Nomor 2, Maret 2026, Hal 465-472
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi



Gambar 4. *Confusion Matrix*

Visualisasi *Confusion matrix* menunjukkan bahwa *Logistic Regression* menghasilkan 149 *true negative* dan 34 *true positive*, dengan 11 *false positive* dan 6 *false negative* pada data uji. Sementara itu, model SVM memberikan hasil yang lebih baik dengan 158 *true negative* dan 36 *true positive*, serta hanya 2 *false positive* dan 4 *false negative*, yang berarti SVM lebih sedikit salah mendeteksi baik pasien non-PCOS maupun PCOS dibanding *Logistic Regression* sehingga lebih unggul untuk keperluan skrining.

3. KESIMPULAN

Penerapan tahapan *data mining* yang meliputi pra-pemrosesan, seleksi fitur dengan RFE, serta pembangunan model dalam bentuk *pipeline* terbukti mampu menghasilkan model prediksi PCOS yang stabil dan dapat diandalkan pada dataset klinis yang digunakan. Pra-pemrosesan berupa pembersihan data, *stratified train-test split* 80:20 dan standarisasi fitur memastikan data berada dalam kondisi bersih dan terukur sehingga proses pembelajaran model berjalan optimal. Seleksi fitur menggunakan RFE berbasis *Logistic Regression* berhasil mengidentifikasi kombinasi atribut klinis yang paling berpengaruh terhadap diagnosis PCOS, yaitu fitur-fitur yang berkaitan dengan status menstruasi, indeks massa tubuh, kadar hormon dan karakteristik folikel. Penggunaan subset fitur terpilih ini tidak hanya menyederhanakan model, tetapi juga membantu meningkatkan performa klasifikasi dan mempermudah interpretasi klinis dibandingkan penggunaan seluruh fitur secara langsung.

Hasil evaluasi menunjukkan bahwa model SVM yang telah dioptimasi *hyperparameter* memberikan kinerja terbaik dibanding *Logistic Regression*, dengan akurasi 0,97, *precision* 0,95, *recall* 0,90, *F1-score* 0,92 dan AUC 0,9977 pada data uji. Nilai-nilai ini lebih tinggi daripada *Logistic Regression* yang menghasilkan akurasi 0,915, *F1-score* 0,80 dan AUC 0,9775, serta c*onfusion matrix* SVM juga mencatat jumlah *false positive* dan *false negative* yang lebih rendah. Temuan ini mengindikasikan bahwa SVM lebih efektif dalam membedakan pasien PCOS dan non-PCOS pada konfigurasi fitur yang digunakan, sehingga lebih direkomendasikan sebagai model utama untuk sistem pendukung keputusan skrining PCOS, sementara *Logistic Regression* tetap bermanfaat sebagai model pembanding yang mudah diinterpretasikan.

UCAPAN TERIMAKASIH

Terima kasih disampaikan kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini.

DAFTAR PUSTAKA

[1] Z. Zad *et al.*, “Predicting *Polycystic Ovary Syndrome* with *machine learning* algorithms from electronic health records,” no. January, pp. 1–14, 2024, doi: 10.3389/fendo.2024.1298628.
[2] S. El-sappagh and H. Saleh, “*Polycystic Ovary Syndrome* Detection *Machine learning* Model Based on Optimized *Feature selection* and Explainable Artificial Intelligence,” pp. 1–21, 2023.
[3] P. Chauhan, “Comparative Analysis of *Machine learning* Algorithms for Prediction of PCOS,” pp. 3–9, 2021.
[4] N. T. Pitaloka, R. Dan, E. D. A. Dengan, and M. Algoritma, “PCOS DISEASE CLASSIFICATION USING *FEATURE SELECTION* RFECV AND KLASIFIKASI PENYAKIT PCOS MENGGUNAKAN *FEATURE SELECTION*,” vol. 4, no. 4, pp. 693–701, 2023.
[5] G. Pandya and D. Solanki, “PREDICTIVE MODELING OF *POLYCYSTIC OVARY SYNDROME* USING *MACHINE LEARNING*,” vol. 20, no. 4, pp. 172–179, 2025.
[6] K. N. Neighbor, “A Comparative Study to Predict Polycystic Ovarian Syndrome (PCOS) Based on Different Models of *Machine learning* Technique,” vol. IV, no. 2, pp. 1–6, 2023.
[7] U. N. Wisesty and T. Mutiah, “Implementasi Gabor Wavelet dan *Support Vector Machine* pada Deteksi Polycystic Ovary (PCO) Berdasarkan Citra,” vol. 1, no. August, pp. 67–82, 2016, doi: 10.21108/indojc.2016.1.2.90.
[8] T. Mahmud and S. A. S. M. Naim, “Predicting *Polycystic Ovary Syndrome* using SVM,” 2024.
[9] P. Divitha, N. S. Vaishnavi, M. Sudeepthi, Y. Sruthi, V. Raghubathy, and S. V. Kumar, “Smart Health Care : *Machine learning* for PCOS,” no. April, 2025.

[10] S. A. Suha and M. N. Islam, "Heliyon Exploring the dominant features and data-driven detection of *Polycystic Ovary Syndrome* through modified stacking ensemble *machine learning* technique," *Heliyon*, vol. 9, no. 3, p. e14518, 2023, doi: 10.1016/j.heliyon.2023.e14518.

[11] M. S. Rohman and T. N. Rahmawati, "Muhammad Syaifur Rohman, 2) Tsalisa Noor Rahmawati 2)," vol. 10, no. 2, pp. 567–575, 2025.

[12] B. Panjwani, J. Yadav, V. Mohan, and N. Agarwal, "Optimized *Machine learning* for the Early Detection of *Polycystic Ovary Syndrome* in Women," 2025.