TESIS

PENINGKATAN KINERJA KLASIFIKASI DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DENGAN PARTICLE SWARM OPTIMIZATION (PSO)

disusun oleh

HASIM AS'ARI

23.55.2524

Konsentrasi : Business Intelegence

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

TESIS

PENINGKATAN KINERJA KLASIFIKASI DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DENGAN PARTICLE SWARM OPTIMIZATION (PSO)

IMPROVING DIABETES CLASSIFICATION PERFORMANCE USING THE SUPPORT VECTOR MACHINE (SVM) METHOD WITH PARTICLE SWARM OPTIMIZATION (PSO)

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi S2 PJJ Informatika

disusun oleh

HASIM AS'ARI

23.55.2524

Konsentrasi : Business Intelegence

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

PENINGKATAN KINERJA KLASIFIKASI DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DENGAN PARTICLE SWARM OPTIMIZATION (PSO)

IMPROVING DIABETES CLASSIFICATION PERFORMANCE USING THE SUPPORT VECTOR MACHINE (SVM) METHOD WITH PARTICLE SWARM OPTIMIZATION (PSO)

yang disusun dan diajukan oleh

Hasim As'ari

23.55.2524

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 3 Februari 2026

Dosen Pembimbing,

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PENGESAHAN

PENINGKATAN KINERJA KLASIFIKASI DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DENGAN PARTICLE SWARM OPTIMIZATION (PSO)

IMPROVING DIABETES CLASSIFICATION PERFORMANCE USING THE SUPPORT VECTOR MACHINE (SVM) METHOD WITH PARTICLE SWARM OPTIMIZATION (PSO)

yang disusun dan diajukan oleh

Hasim As'ari

23.55.2524

Telah dipertahankan di depan Dewan Penguji
pada tanggal 3 Februari 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Dr. Ferry Wahyu Wibowo, S.St., M.Cs. **NIK. 190302235**	
Dr. Sri Ngudi Wahyuni, ST., M.Kom **NIK. 190302060**	
Prof. Dr. Kusrini, M.Kom. **NIK. 190302106**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 3 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Hasim As'ari
NIM : 23.55.2524
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Peningkatan Kinerja Klasifikasi Diabetes Menggunakan Metode Support Vector Machine (SVM) dengan Particle Swarm Optimization (PSO)

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian PENELITI sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab PENELITI, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini PENELITI buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka PENELITI bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 3 Februari 2026
Yang Menyatakan,

Hasim As'ari

HALAMAN PERSEMBAHAN

Dengan mengucapkan Alhamdulillah serta dengan penuh rasa syukur, tesis ini dipersembahkan kepada orang tua dan istri yang selalu memberikan doa, dukungan, dan kasih peneliting tanpa batas, serta kepada keluarga, dosen pembimbing, dan rekan-rekan yang telah memberikan motivasi, ilmu, serta bimbingan dalam menyelesaikan penelitian ini. Semoga karya ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan serta menjadi langkah awal dalam kontribusi nyata di bidang penelitian dan teknologi.

HALAMAN MOTTO

"Dalam setiap usaha terdapat ikhtiar manusia, dan dalam setiap hasil terdapat ketetapan Allah. Ilmu yang dipelajari adalah bentuk ikhtiar, sedangkan keberhasilan adalah karunia-Nya. Karena itu, manusia wajib berusaha sebaik mungkin, memohon petunjuk-Nya, dan berserah diri atas hasil akhirnya. Dengan memadukan ilmu, kerja keras, dan tawakal, setiap langkah akan menjadi lebih bermakna.

KATA PENGANTAR

Puji syukur kehadirat Allah Ta'ala atas limpahan rahmat, taufik, dan hidayah-Nya, sehingga penulis dapat menyelesaikan tesis yang berjudul "Peningkatan Kinerja Klasifikasi Diabetes Menggunakan Metode Support Vector Machine (SVM) dengan Particle Swarm Optimization (PSO)". Melalui proses penelitian dan penyusunan tesis ini, penulis memperoleh banyak ilmu, pengalaman, serta wawasan baru, khususnya dalam bidang machine learning dan analisis data medis.

Penyusunan tesis ini tidak terlepas dari dukungan, bantuan, dan bimbingan berbagai pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada dosen pembimbing yang telah dengan sabar memberikan arahan, masukan, serta motivasi yang sangat berarti selama proses penelitian. Ucapan terima kasih juga penulis tujukan kepada dosen penguji atas kritik dan saran yang membangun demi penyempurnaan penelitian ini.

Tak lupa, rasa terima kasih yang tulus penulis sampaikan kepada keluarga tercinta, teman-teman, serta seluruh pihak yang telah memberikan dukungan moral maupun materiil selama penyusunan tesis ini. Semangat dan doa mereka menjadi kekuatan bagi penulis dalam menyelesaikan penelitian ini.

Akhir kata, penulis berharap penelitian ini dapat memberikan kontribusi bagi pengembangan ilmu pengetahuan, khususnya dalam bidang klasifikasi penyakit berbasis machine learning, serta menjadi referensi bagi penelitian selanjutnya. Penulis juga menyadari bahwa tesis ini masih memiliki keterbatasan,

sehingga kritik dan saran yang membangun sangat diharapkan demi kesempurnaan penelitian di masa mendatang. Semoga hasil penelitian ini dapat memberikan manfaat bagi banyak pihak.

Yogyakarta, tanggal bulan tahun

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Diabetes mellitus merupakan penyakit kronis yang memerlukan deteksi dini untuk mengurangi risiko komplikasi. Penelitian ini bertujuan mengembangkan model klasifikasi diabetes menggunakan Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) yang dioptimasi menggunakan Particle Swarm Optimization (PSO). Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang memiliki distribusi kelas tidak seimbang. Penelitian ini menerapkan pipeline yang terdiri dari StandardScaler, SMOTE, SelectKBest, dan SVM-RBF. PSO digunakan untuk mengoptimasi parameter SVM serta jumlah fitur terbaik, sementara optimasi threshold dilakukan untuk meningkatkan keseimbangan antara precision dan recall. Hasil pengujian menunjukkan optimal diperoleh pada C = 20, gamma = 0,00699, dengan tujuh fitur terpilih. Model menghasilkan akurasi sebesar 77,6%, recall 80,6%, precision 64,3%, F1-score 71,5%, dan ROC-AUC 83,6%. Hasil tersebut menunjukkan bahwa model SVM-RBF yang dioptimasi menggunakan PSO mampu mendeteksi pasien diabetes dengan baik dan memiliki potensi sebagai sistem pendukung keputusan medis untuk skrining awal diabetes. otomatis.

Kata Kunci : Diabetes Mellitus, SVM, PSO, SMOTE, Klasifikasi.

ABSTRACT

Diabetes mellitus is a chronic disease that requires early detection to reduce the risk of complications. This study aims to develop a diabetes classification model using a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel optimized by Particle Swarm Optimization (PSO). The Pima Indians Diabetes Dataset, which exhibits class imbalance, was used in this study. An integrated modeling pipeline consisting of StandardScaler, SMOTE, SelectKBest, and SVM-RBF was implemented. PSO was applied to optimize SVM parameters and select the optimal number of features, while threshold optimization was performed to balance precision and recall. The optimal configuration achieved C = 20, gamma = 0.00699, with seven selected features. The model achieved an accuracy of 77.6%, recall of 80.6%, precision of 64.3%, F1-score of 71.5%, and ROC-AUC of 83.6%. These results indicate that the PSO-optimized SVM-RBF model performs well in detecting diabetic patients and has strong potential as a medical decision support system for early diabetes screening.

Keywords: Diabetes Mellitus, SVM, PSO, SMOTE, Classification.

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Diabetes melitus merupakan salah satu penyakit metabolik kronis yang menjadi masalah kesehatan global dengan prevalensi yang terus meningkat. Penyakit ini terjadi karena ketidakmampuan tubuh untuk memproduksi atau menggunakan insulin secara efektif, yang mengakibatkan kadar gula darah tinggi. Berdasarkan laporan Organisasi Kesehatan Dunia (WHO), lebih dari 422 juta orang di dunia menderita diabetes pada tahun 2021, dengan angka kematian akibat komplikasi penyakit ini mencapai 1,5 juta per tahun (World Health Organization, 2021). Diabetes tidak hanya berdampak pada kesehatan individu, tetapi juga menjadi beban sosial dan ekonomi yang signifikan, terutama di negara-negara berkembang.

Penyakit kronis ini juga ditandai dengan tingginya kadar gula darah akibat gangguan produksi atau fungsi insulin. Peningkatan prevalensi diabetes secara global menuntut pengembangan metode diagnostik yang lebih akurat dan efisien. Perlu ada upaya proaktif untuk menghindari timbulnya diabetes. Hal ini mencakup berbagai langkah seperti pola hidup sehat (Wiardani & Kusumajaya, 2023), vaksinasi, pemeriksaan kesehatan rutin, serta diketahuinya penyakit tersebut. Salah satu pendekatan yang digunakan dalam klasifikasi diabetes adalah penerapan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF). SVM dengan kernel RBF mampu menangani data dengan pola yang

kompleks dan non-linear, sehingga cocok untuk aplikasi medis seperti prediksi diabetes.

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin berbasis kernel yang dirancang untuk menyelesaikan tugas klasifikasi dengan membangun hyperplane guna memisahkan data ke dalam kelas-kelas yang berbeda (Javeed et al., 2023). Algoritma ini berfungsi dengan mencari hyperplane atau garis pemisah terbaik yang mampu memisahkan data dari berbagai kelas secara maksimal. Pada proses klasifikasi, SVM menggunakan data yang paling dekat dengan hyperplane, yang dikenal sebagai *support vectors*, untuk menentukan posisi optimal dari garis pemisah tersebut.

Penelitian oleh (Arora et al., 2022) berfokus pada pengembangan model prediktif untuk klasifikasi diabetes dengan menggabungkan algoritma *K-means clustering* dan *Support Vector Machine* (SVM). Dalam pendekatan ini, algoritma *K-means* digunakan untuk mengelompokkan data menjadi cluster yang lebih homogen, sehingga mempermudah proses klasifikasi oleh SVM. Dataset yang digunakan adalah Pima Indians Diabetes, yang terdiri dari delapan fitur independen seperti kadar glukosa, tekanan darah, dan indeks massa tubuh, serta satu variabel dependen untuk diagnosis diabetes. Dengan memanfaatkan data dari 668 pasien perempuan, model ini melibatkan 80% data untuk pelatihan dan 20% untuk pengujian. Hasil penelitian menunjukkan akurasi sebesar 98,7%, yang jauh lebih tinggi dibandingkan metode klasifikasi lainnya, seperti *decision tree* dan *naive Bayes*. Selain akurasi, penelitian ini juga mengevaluasi indikator lain, seperti presisi, recall, dan F1-score, untuk mengukur performa model secara keseluruhan,

menunjukkan bahwa kombinasi *K-means* dan SVM mampu meningkatkan akurasi prediksi dengan signifikan

Beberapa penelitian oleh (Shrestha et al., 2023) mengusulkan solusi baru untuk meningkatkan prediksi onset Diabetes Mellitus Tipe 2 dengan menggunakan kombinasi teknik pembelajaran mendalam (deep learning), yaitu algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) dan lapisan Long Short-Term Memory (LSTM). Dalam penelitian ini, RBF digunakan untuk mengoptimalkan akurasi klasifikasi, sementara LSTM membantu menangkap pola dalam data yang memiliki ketergantungan temporal. Hasilnya menunjukkan peningkatan akurasi hingga 86,31% dengan nilai AUC rata-rata sebesar 82,70%, yang lebih baik dibandingkan metode standar industri. Selain itu, waktu pemrosesan berhasil ditingkatkan sebesar 3,8 milidetik, membuat pendekatan ini lebih efisien untuk aplikasi praktis. Dengan memanfaatkan dataset seperti Pima Indians dan Global Diabetes Health Record, penelitian ini memberikan kontribusi signifikan dalam meningkatkan keakuratan dan efisiensi prediksi Diabetes Mellitus Tipe 2, sekaligus mengurangi risiko kesalahan prediksi dan waktu pemrosesan

Selain itu, Penelitian yang dilakukan oleh (Lumbanraja et al., 2022) bertujuan untuk mengembangkan model prediksi klasifikasi penderita Diabetes Mellitus menggunakan algoritma Support Vector Machine (SVM). Dataset yang digunakan berasal dari *Diabetes 130-US Hospitals for Years 1999-2008*, dengan jumlah data awal sebanyak 101,766 rekaman yang setelah proses pembersihan menjadi 84,900 data. Penelitian ini menerapkan metode *10-fold cross-validation* dan membandingkan tiga jenis kernel pada SVM, yaitu linear, Gaussian, dan

polynomial. Hasil penelitian menunjukkan bahwa kernel Gaussian memberikan akurasi tertinggi sebesar 82,76%, sedangkan kernel linear dan polynomial masing-masing menghasilkan akurasi sebesar 72,48% dan 39,56%. Penelitian ini juga mengidentifikasi bahwa ketidakseimbangan data (imbalance) memengaruhi performa klasifikasi, terutama pada nilai sensitivitas. Oleh karena itu, pengelolaan data yang lebih baik direkomendasikan untuk penelitian selanjutnya guna meningkatkan sensitivitas dan spesifisitas model.

Penelitian ini berfokus pada penerapan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) untuk klasifikasi diabetes berdasarkan dataset yang diperoleh dari Kaggle, yang mencakup 769 data pasien. Data tersebut terdiri dari berbagai parameter kesehatan seperti jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, berat badan, faktor keturunan, dan usia. Penelitian ini melibatkan tahap preprocessing, termasuk normalisasi data untuk mengatasi masalah seperti nilai yang hilang dan format yang tidak sesuai. Dataset kemudian dibagi menjadi data training dan testing dengan berbagai skenario pembagian, seperti 80:20, 70:30, dan 90:10. Hasil menunjukkan bahwa algoritma SVM dengan kernel RBF mencapai akurasi tertinggi sebesar 87% pada skenario tertentu, menunjukkan kemampuan algoritma ini dalam menangkap pola non-linear yang kompleks dalam data kesehatan. Hal ini menjadikan metode ini efektif untuk mendukung deteksi dini penyakit diabetes.(Muhammad Hilmy Haidar Aly, 2024)

Dataset diabetes yang peneliti gunakan memiliki fitur Tekanan Darah dan Umur dengan rentang nilai yang tidak dalam skala tertentu, sehingga perlu

dilakukan normalisasi agar skala data lebih seragam. Peneliti menggunakan Min-Max Scaler untuk mengubah nilai-nilai dalam dataset ke dalam rentang 0 hingga 1, sehingga semua fitur memiliki skala yang sama tanpa mengubah distribusi datanya. Tujuan utama penggunaan Min-Max Scaler adalah untuk meningkatkan performa model dengan memastikan bahwa fitur dengan rentang nilai yang besar tidak mendominasi proses pembelajaran Algoritma ini (He, 2013a). Setelah Min-Max Scaler diterapkan, nilai Tekanan Darah dan Umur menjadi lebih terdistribusi secara proporsional dalam skala 0 hingga 1, yang membantu algoritma dalam mengoptimalkan proses klasifikasi tanpa bias terhadap fitur tertentu.

Meskipun demikian, tantangan utama dalam penerapan SVM-RBF pada klasifikasi Diabetes Mellitus terletak pada kualitas dan karakteristik dataset. Dataset diabetes yang banyak digunakan, seperti Pima Indians Diabetes Dataset, memiliki beberapa permasalahan utama, yaitu perbedaan skala antar fitur, jumlah data yang relatif terbatas, serta distribusi kelas yang tidak seimbang antara pasien diabetes dan non-diabetes (Reza et al., 2023a). Ketidakseimbangan kelas ini berpotensi menyebabkan model bias terhadap kelas mayoritas, sehingga meskipun nilai akurasi terlihat tinggi, kemampuan model dalam mendeteksi pasien diabetes (kelas minoritas) menjadi kurang optima (Salmi et al., 2024).

Meskipun berbagai algoritma machine learning seperti Logistic Regression, Decision Tree, Random Forest, dan Neural Network telah digunakan dalam klasifikasi Diabetes Mellitus, pemilihan algoritma yang tepat harus mempertimbangkan karakteristik data medis yang umumnya bersifat non-linear, berdimensi terbatas, serta tidak seimbang (Afolabi et al., 2025). Pada kondisi

tersebut, Support Vector Machine (SVM) menjadi pilihan yang relevan karena kemampuannya dalam membangun batas keputusan optimal berbasis margin maksimum dan ketahanannya terhadap overfitting.

Kernel Radial Basis Function (RBF) dipilih karena kemampuannya dalam memetakan data non-linear ke ruang berdimensi lebih tinggi tanpa meningkatkan kompleksitas komputasi secara signifikan (Du et al., 2024). Dibandingkan kernel linear, kernel RBF lebih fleksibel dalam menangkap hubungan kompleks antar variabel medis seperti kadar glukosa, indeks massa tubuh, dan usia. Sementara itu, dibandingkan model ensemble atau deep learning, SVM-RBF relatif lebih stabil pada dataset berukuran terbatas dan tidak memerlukan sumber daya komputasi yang besar, sehingga lebih sesuai untuk implementasi praktis di bidang kesehatan.

Pemanfaatan teknik machine learning, khususnya Support Vector Machine (SVM), telah banyak digunakan dalam klasifikasi penyakit diabetes karena kemampuannya dalam menangani data berdimensi tinggi dan pola nonlinier. Namun, kinerja SVM sangat bergantung pada pemilihan parameter kernel dan fitur yang digunakan, sehingga hasil yang tidak optimal dapat menurunkan akurasi dan kemampuan generalisasi model. Oleh karena itu, diperlukan metode optimasi yang efektif untuk meningkatkan performa SVM, salah satunya adalah Particle Swarm Optimization (PSO). PSO memiliki keunggulan dalam mencari solusi optimal secara global melalui mekanisme pencarian berbasis populasi yang sederhana dan efisien. Dengan mengombinasikan SVM dan PSO, diharapkan dapat diperoleh model klasifikasi diabetes yang memiliki kinerja lebih baik, stabil, dan mampu mendukung proses skrining serta deteksi dini diabetes secara lebih akurat.optimal.

Particle Swarm Optimization (PSO) merupakan algoritma optimasi berbasis populasi yang terinspirasi dari perilaku sosial kawanan, seperti burung atau ikan, dalam mencari sumber makanan. Dalam PSO, setiap solusi direpresentasikan sebagai partikel yang bergerak di dalam ruang pencarian dengan kecepatan tertentu. Pergerakan partikel dipengaruhi oleh dua informasi utama, yaitu pengalaman terbaik partikel itu sendiri (personal best/pbest) dan pengalaman terbaik seluruh populasi (global best/gbest). Melalui mekanisme ini, partikel secara kolektif mengeksplorasi ruang solusi dan secara bertahap bergerak menuju solusi optimal (Kennedy & Eberhart, 2022).

Berbagai algoritma machine learning telah banyak diterapkan dalam klasifikasi penyakit berbasis data medis, seperti Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Neural Network, serta Support Vector Machine (SVM). Namun, efektivitas masing-masing metode sangat bergantung pada karakteristik data yang digunakan. Data medis umumnya memiliki hubungan antar variabel yang bersifat non-linear, jumlah data yang relatif terbatas, serta distribusi kelas yang tidak seimbang, sehingga pemilihan metode klasifikasi harus mempertimbangkan aspek akurasi, stabilitas, dan kemampuan generalisasi model (Afolabi et al., 2025).

Logistic Regression merupakan metode klasifikasi linier yang sederhana dan mudah diinterpretasikan, tetapi memiliki keterbatasan dalam memodelkan hubungan non-linear antar variabel klinis (Kuhn & Johnson, 2013). KNN mengandalkan kedekatan jarak antar data, namun sensitif terhadap skala fitur dan noise, serta kurang efisien pada data berdimensi tinggi (Géron, 2019). Decision

Tree mampu menangani hubungan non-linear, tetapi cenderung mengalami overfitting pada dataset berukuran kecil (Han & Kamber, 2012). Sementara itu, Random Forest dan Neural Network mampu menghasilkan akurasi yang tinggi, namun memiliki kompleksitas komputasi yang lebih besar serta interpretabilitas yang lebih rendah, sehingga kurang optimal untuk implementasi praktis dalam sistem pendukung keputusan medis (Junus et al., 2023).

Berdasarkan keterbatasan tersebut, Support Vector Machine (SVM) dipilih sebagai metode klasifikasi utama dalam penelitian ini. SVM memiliki kemampuan untuk membangun hyperplane optimal dengan margin maksimum, sehingga lebih stabil dan memiliki kemampuan generalisasi yang baik pada dataset berukuran terbatas (Kuhn & Johnson, 2013). Penggunaan kernel Radial Basis Function (RBF) memungkinkan SVM menangani pola non-linear yang kompleks tanpa meningkatkan kompleksitas komputasi secara signifikan (Du et al., 2024). Namun, kinerja SVM sangat dipengaruhi oleh pemilihan parameter C dan gamma, sehingga diperlukan metode optimasi yang efektif. Oleh karena itu, penelitian ini mengombinasikan SVM dengan Particle Swarm Optimization (PSO) untuk memperoleh konfigurasi parameter dan fitur yang optimal, sehingga mampu meningkatkan kinerja klasifikasi diabetes secara lebih stabil dan sensitif terhadap kelas minoritas

Penelitian ini menggunakan PIMA Indian Diabetes Dataset dan menerapkan beberapa algoritma klasifikasi, yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest, untuk memprediksi diabetes pada tahap awal. Hasil penelitian menunjukkan bahwa algoritma Random Forest

memberikan kinerja terbaik dengan tingkat akurasi tertinggi sebesar 83,11%, dibandingkan SVM dengan akurasi maksimal 75%, KNN sekitar 79,78%, dan Decision Tree sebesar 79,68%. Temuan ini menunjukkan bahwa pendekatan ensemble learning, khususnya Random Forest, lebih efektif dalam memodelkan data diabetes dan berpotensi besar untuk diterapkan sebagai alat bantu pengambilan keputusan dalam sistem pelayanan kesehatan. (Kumar, 2022)

Berdasarkan penelitian sebelumnya oleh Kumar (2022), klasifikasi diabetes umumnya dilakukan dengan membandingkan beberapa algoritma machine learning, di mana Random Forest menunjukkan akurasi tertinggi. Namun, penelitian tersebut belum mengkaji optimasi hyperparameter secara terintegrasi, penanganan ketidakseimbangan data dalam pipeline yang bebas kebocoran data (leakage-free), serta optimasi threshold untuk meningkatkan sensitivitas terhadap kelas minoritas. Oleh karena itu, kebaruan penelitian ini terletak pada pengembangan pipeline klasifikasi diabetes berbasis Support Vector Machine dengan kernel Radial Basis Function (SVM-RBF) yang dioptimasi menggunakan Particle Swarm Optimization (PSO), dikombinasikan dengan SMOTE, seleksi fitur, dan optimasi threshold, sehingga menghasilkan model yang lebih stabil, sensitif, dan memiliki kemampuan generalisasi yang lebih baik.

Mempertimbangkan karakteristik dataset diabetes yang digunakan, serta kebutuhan akan model yang akurat, stabil, dan memiliki kemampuan generalisasi yang baik, penelitian ini secara khusus memfokuskan penggunaan algoritma Support Vector Machine dengan kernel Radial Basis Function (SVM-RBF). Dengan pendekatan tersebut, penelitian ini diharapkan tidak hanya meningkatkan

nilai akurasi klasifikasi dalam mendeteksi kasus Diabetes Mellitus, khususnya pada kelas minoritas. Hasil penelitian ini diharapkan dapat memberikan kontribusi metodologis dalam pengembangan sistem pendukung keputusan berbasis machine learning di bidang kesehatan.

1.2. Rumusan Masalah

Berdasarkan uraian pada latar belakang masalah, rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

a. Bagaimana pengaruh pipeline preprocessing yang terdiri dari StandardScaler, SelectKBest, dan SMOTE terhadap kinerja klasifikasi Diabetes Mellitus menggunakan algoritma SVM-RBF?
b. Berapakah nilai optimal hyperparameter C dan gamma (γ) pada SVM-RBF yang mampu menghasilkan kinerja klasifikasi terbaik?

1.3. Batasan Masalah

Berdasarkan penjelasan pada latar belakang masalah, batasan masalah dalam penelitian ini dirumuskan sebagai berikut:

a. Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang diperoleh dari platform Kaggle.
b. Penelitian difokuskan pada klasifikasi Diabetes Mellitus ke dalam dua kelas, yaitu diabetes dan tidak diabetes.
c. Algoritma yang digunakan adalah Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF).

d. Tahapan preprocessing meliputi StandardScaler, SelectKBest, dan SMOTE.

e. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC.

1.4. Tujuan Penelitian

Berdasarkan uraian dalam latar belakang masalah, tujuan dari penelitian ini dirumuskan sebagai berikut:

a. Menganalisis kinerja algoritma SVM-RBF dalam mengklasifikasikan Diabetes Mellitus menggunakan pipeline preprocessing.

b. Menentukan nilai optimal hyperparameter C dan gamma (γ) pada SVM-RBF untuk memperoleh performa klasifikasi terbaik.

1.5. Manfaat Penelitian

Berdasarkan penjelasan dalam latar belakang masalah, manfaat dari penelitian ini adalah sebagai berikut:

1.5.1 Manfaat Akademis

Memberikan kontribusi ilmiah dalam pengembangan metode klasifikasi Diabetes Mellitus berbasis machine learning, khususnya dalam optimalisasi kinerja SVM-RBF melalui preprocessing terintegrasi.

1.5.2 Manfaat Praktis

Menjadi referensi dalam pengembangan sistem pendukung keputusan untuk deteksi dini Diabetes Mellitus yang akurat dan efisien.

1.5.3 Manfaat Penelitian Lanjutan

Menjadi acuan bagi penelitian selanjutnya dalam penerapan dan pengembangan algoritma machine learning pada bidang kesehatan dengan karakteristik data yang tidak seimbang dan non-linear.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Dalam penelitian ini, penulis memanfaatkan informasi dari berbagai penelitian terdahulu sebagai bahan perbandingan untuk mengevaluasi kelebihan dan kekurangannya. Langkah ini bertujuan untuk memahami keterkaitan antara penelitian sebelumnya dengan penelitian yang sedang dilakukan. Oleh karena itu, studi-studi terdahulu memiliki peranan yang krusial bagi pengembangan penelitian ini. Selain itu, tinjauan literatur dilakukan untuk menegaskan bahwa penelitian ini memberikan manfaat dan kontribusi yang berarti bagi ilmu pengetahuan.

Berbekal ketersediaan dataset Diabetes Mellitus yang dapat diakses secara publik, penulis terdorong untuk mengembangkan penelitian lebih lanjut dengan menggunakan dataset tersebut untuk mengidentifikasi penyakit Diabetes Mellitus berdasarkan data terkait gaya hidup. Dalam implementasi algoritma Support Vector Machine (SVM), penulis mengacu pada pendekatan yang telah dianalisis dalam penelitian sebelumnya. Berikut ini adalah beberapa ulasan mengenai penerapan algoritma SVM.

Pada penelitian (Arora et al., 2022) memprediksi pasien diabetes menggunakan teknik pengelompokan K-means dan Support Vector Machine (SVM). Hal ini untuk menciptakan model prediksi dengan tingkat accuracy tinggi yang dapat mengidentifikasi kemungkinan seseorang menderita diabetes. Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang berisi data 668 pasien

wanita. Support Vector Machine (SVM) berperan sebagai algoritma utama untuk klasifikasi data diabetes setelah melalui proses pengelompokan dengan K-means. Algoritma ini bekerja dengan menentukan hyperplane yang memaksimalkan margin antara data dari kelas yang berbeda, sehingga meningkatkan nilai kurasi. Kombinasi SVM dengan K-means memungkinkan pemrosesan yang lebih terstruktur, di mana data yang telah dikelompokkan menjadi fitur yang lebih representatif untuk SVM dalam proses klasifikasi. Hasil penelitian menunjukkan bahwa pendekatan kombinasi K-means dan SVM memberikan kinerja yang lebih baik. Pendekatan ini mencapai accuracy sebesar 98,7%, lebih tinggi dibandingkan hanya menggunakan SVM yang mencapai 82,46%. Selain itu, hasil metrik lainnya seperti Pressision (98,6%), Recall (96,8%), dan F1-score (97,5%) juga menunjukkan keunggulan dari metode ini. Tingginya tingkat accuracy ini membuktikan bahwa pendekatan dengan dua tahap tersebut dapat mengatasi kelemahan klasifikasi yang menggunakan satu algoritma saja.

Pada penelitian (Shrestha et al., 2023) Hasil penelitian ini secara sistematis menunjukkan keberhasilan metode prediksi Diabetes Mellitus tipe 2 dengan mengintegrasikan algoritma Support Vector Machine (SVM) menggunakan kernel Radial Basis Function (RBF) dan Long Short-Term Memory (LSTM). Proses penelitian mencakup tahapan ekstraksi fitur, analisis, dan validasi menggunakan dataset publik seperti Pima Indians Diabetes dan Global Diabetes Health Record. Hasilnya, metode ini berhasil meningkatkan akurasi prediksi hingga 86,31% dan nilai Area Under the Curve (AUC) menjadi 82,70%, melampaui standar industri. Selain itu, waktu pemrosesan berhasil dipangkas hingga 3,8 milidetik,

meningkatkan efisiensi untuk aplikasi praktis. Kelebihan utama dari pendekatan ini adalah kemampuan untuk menangani dataset kompleks dan meningkatkan performa prediksi secara signifikan melalui pengelolaan dependensi data jangka panjang oleh LSTM. Namun, kelemahan dari metode ini adalah kebutuhan komputasi yang tinggi, yang memerlukan perangkat keras dengan spesifikasi tinggi, serta keterbatasan pengujian pada dataset yang lebih besar dan beragam. Selain itu, penelitian ini kurang menyoroti metrik penting lainnya seperti sensitivitas dan spesifisitas, yang relevan dalam konteks aplikasi medis untuk mengurangi kesalahan diagnosis.

Penelitian yang lain (Lumbanraja et al., 2022) Penelitian ini menganalisis implementasi Support Vector Machine (SVM) untuk klasifikasi penderita diabetes mellitus menggunakan dataset dari UCI Machine Learning Repository. Data terdiri dari 34 variabel dan 84.900 entri, dengan teknik validasi menggunakan 10-fold cross-validation serta tiga kernel SVM: linear, Gaussian, dan polynomial. Hasilnya menunjukkan kernel Gaussian memiliki akurasi tertinggi sebesar 82,76%, dibandingkan kernel linear (72,48%) dan polynomial (39,56%). Kelebihan metode ini adalah akurasi tinggi pada kernel Gaussian dan kemampuan menangani data kompleks. Namun, kelemahan utama adalah sensitivitas rendah akibat ketidakseimbangan data, di mana jumlah data negatif jauh lebih besar dibanding data positif, sehingga memengaruhi performa klasifikasi. Implementasi sistem berbasis R Shiny juga mempermudah pengguna dalam menganalisis prediksi, meskipun peningkatan pada pengelolaan data imbang tetap diperlukan

Penelitian yang dilakukan oleh Hilmy (Muhammad Hilmy Haidar Aly, 2024) Hasil penelitian menunjukkan bahwa algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) berhasil mencapai akurasi maksimal sebesar 87% dalam klasifikasi diabetes, menggunakan dataset dari Kaggle yang berisi 769 catatan. Proses dimulai dari preprocessing data untuk menghilangkan masalah seperti nilai yang hilang dan skala data, diikuti oleh pembagian data menjadi training dan testing, dan pemodelan menggunakan SVM dengan kernel RBF. Kelebihan penelitian ini adalah kemampuan SVM-RBF untuk menangkap hubungan nonlinier antara atribut kesehatan, sehingga meningkatkan performa klasifikasi. Namun, penelitian ini memiliki kelemahan, seperti ketergantungan pada optimasi parameter gamma dan C untuk hasil terbaik, serta ukuran dataset yang terbatas sehingga perlu divalidasi lebih lanjut menggunakan dataset eksternal untuk memastikan generalisasi model.

Penelitian lainnya yang dilakukan oleh Reza (Reza et al., 2023b) Hasil penelitian ini menunjukkan bahwa kombinasi kernel non-linear baru, yaitu kernel terintegrasi dari Radial Basis Function (RBF) dan RBF City Block, berhasil meningkatkan kinerja klasifikasi diabetes tipe II menggunakan dataset PIMA. Dengan pendekatan ini, akurasi model mencapai 85,5%, recall 83,4%, presisi 87,0%, dan F1-score 85,2%. Penelitian juga menggunakan metode preprocessing seperti imputasi data yang hilang dengan nilai median, penghapusan outlier menggunakan metode IQR, dan penyeimbangan kelas dengan SMOTE. Kelebihan pendekatan ini adalah kemampuan menangani kompleksitas data non-linear, mengurangi overfitting, dan meningkatkan akurasi prediksi. Namun, kelemahan

utamanya adalah kompleksitas computational yang lebih tinggi dibandingkan dengan kernel tunggal dan kebutuhan optimasi parameter yang tepat untuk mencapai kinerja maksimal.

Sedangkan pada penelitian lainnya (Junus et al., 2023) Penelitian ini membandingkan dua metode klasifikasi, yaitu Support Vector Machine (SVM) dan Random Forest (RF), untuk deteksi awal risiko Diabetes Melitus menggunakan data dari UCI Machine Learning Repository. Hasil penelitian menunjukkan bahwa metode Random Forest memiliki performa yang lebih baik dibandingkan SVM, dengan akurasi 98,08%, recall 97,87%, precision 98,92%, dan F1_Score 98,40%. Sebaliknya, SVM menghasilkan akurasi 91,03%, recall 86,05%, precision 97,37%, dan F1_Score 91,36%. Kelebihan Random Forest adalah kemampuannya menangani data dalam jumlah besar dan mengatasi data yang tidak lengkap, serta menghasilkan klasifikasi yang lebih akurat. Namun, kelemahannya adalah kompleksitas model yang lebih tinggi dan waktu komputasi yang lebih lama. Sementara itu, SVM memiliki kelebihan dalam menghasilkan hyperplane terbaik untuk pemisahan kelas data dan akurasi yang tinggi pada data linier, tetapi kurang efektif pada data non-linier dan memerlukan tuning parameter

2.2. Keaslian Penelitian

Tabel 2.2 Matriks Literatur Review dan Posisi penelitian
Pemodelan Prediksi Diabetus Militus Menggunakan Algoritma *Support Vector Machine*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Improving SVM Performance for Type II Diabetes Prediction with an Improved Non-Linear Kernel Insights from The PIMA Dataset	Md. Shamim Reza, Umme Hafsha, Ruhul Amin, Rubia Yasmin dan Sabba Ruhi Computer Methods and Programs in Biomedicine Update, 2023	Meningkatkan performa algoritma *Support Vector Machine* (SVM) dalam memprediksi diabetes tipe II dengan mengembangkan kernel *non-linear* yang lebih baik, menggunakan kombinasi kernel *Radial Basis Function* (RBF) dan *RBF City Block*.	Model SVM dengan kernel terintegrasi yang diusulkan mencapai *accuracy* sebesar 85,5%, *Recall* 87,0%, *Precision* 83,4%, dan *F1-score* 85,2%. Selain itu, AUC juga mencapai 85,5%, yang menunjukkan kemampuan model dalam membedakan antara pasien diabetes dan non-diabetes	Penelitian lebih lanjut diperlukan untuk memvalidasi kinerja kernel non-linear yang ditingkatkan pada beragam dataset dan di seluruh populasi yang berbeda.	Penelitian ini meningkatkan kinerja SVM melalui optimasi parameter dan seleksi fitur menggunakan PSO dalam pipeline terintegrasi. Hasil menunjukkan model stabil dengan ROC-AUC 83,6% dan recall 80,6%.
2	A Novel Architecture for Diabetes Patients' Prediction Using K-Means Clustering and SVM	Nitin Arora, Anupam Singh, Mustafa Zuhaer Nayef Al-Dabagh, dan Sumit Kumar Maitra Mathematical Problems in Engineering, 2022	Mengembangkan arsitektur baru yang dapat memprediksi kemungkinan seseorang terkena diabetes dengan tingkat akurasi yang tinggi. Penelitian ini menggunakan teknik K-Means Clustering yang dikombinasikan dengan algoritma Support Vector Machine (SVM)	Berhasil mengembangkan arsitektur baru untuk prediksi diabetes menggunakan kombinasi K-Means Clustering dan Support Vector Machine (SVM), dengan akurasi mencapai 75% pada Pima Indians Diabetes Database. Pendekatan ini secara signifikan meningkatkan kinerja	Validasi hanya dilakukan pada satu dataset publik, yaitu Pima Indians Diabetes Database, sehingga generalisasi hasilnya terhadap populasi yang lebih luas belum teruji. Selain itu, meskipun akurasinya tinggi, penelitian ini tidak membahas secara mendalam interpretabilitas	Berbeda dengan pendekatan clustering, penelitian ini mengoptimasi SVM menggunakan PSO, seleksi fitur, dan SMOTE. Hasil menunjukkan peningkatan kemampuan deteksi diabetes melalui recall dan ROC-AUC.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				dibandingkan metode lain seperti SVM tunggal (82,46%).	model, yang penting untuk aplikasi medis.	
3	A Novel Solution of Deep Learning for Enhanced Support Vector Machine for Predicting the Onset of Type 2 Diabetes	Marmik Shrestha, Omar Hisham Alsadoon, Abeer Alsadoon, Thair Al-Dala'in, Tarik A. Rashid, P.W.C. Prasad, Ahmad Alrubaie Multimedia Tools and Applications Tahun: 2022	Meningkatkan akurasi dan Area Under the Curve (AUC) dalam memprediksi diabetes tipe 2 serta mempercepat waktu pemrosesan dengan mengusulkan teknik pembelajaran mendalam yang menggabungkan algoritma Support Vector Machine (SVM), Radial Base Function (RBF), dan Long Short-Term Memory (LSTM).	Metode yang diusulkan berhasil meningkatkan akurasi prediksi diabetes hingga rata-rata 86,31% dan AUC 82,70%, dengan pengurangan waktu pemrosesan sebesar 3,8 milidetik dibandingkan standar industri. Kombinasi RBF kernel dan LSTM meningkatkan akurasi dan efisiensi tanpa mengorbankan waktu pemrosesan.	Model SVM tidak menerapkan *hyperparameter tuning*, sehingga potensi peningkatan akurasi dan kinerja model tidak dioptimalkan secara maksimal.	Penelitian ini tidak menggunakan deep learning, tetapi memaksimalkan SVM-RBF melalui PSO. Pendekatan ini menghasilkan kinerja kompetitif dengan kompleksitas dan beban komputasi yang lebih rendah.
4	Implementasi Support Vector Machine untuk Klasifikasi Penderita Diabetes Mellitus	Favorisen Rosyking Lumbanraja, Fanni Lufiana, Yunda Heningtyas, dkk.,	Mengamplementasikan algoritma Support Vector Machine dalam mengklasifikasikan penderita diabetes mellitus dan membandingkan kinerjanya dengan algoritma Naive Bayes	SVM dengan kernel RBF menghasilkan akurasi tertinggi (97,3%) dalam klasifikasi diabetes mellitus, dibandingkan dengan Naive Bayes (76,3%) dan KNN (96,3%) (Lumbanraja et al., 2022). Hal ini	Dataset tidak ditangani dengan teknik *imbalanced data* maupun *Min-Max Scaler*, sehingga distribusi data kurang optimal. Selain itu, model tidak menerapkan *hyperparameter tuning*,	Penelitian ini menambahkan penanganan data tidak seimbang dan optimasi hyperparameter SVM. Hasil menunjukkan kemampuan generalisasi dan evaluasi model yang lebih baik.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Ilmu Komputer Unila Publishing Network pada tahun 2022	dan K-Nearest Neighbor dan untuk mengevaluasi efektivitas SVM dalam mendiagnosis diabetes berdasarkan dataset yang tersedia.	menunjukkan bahwa SVM berpotensi menjadi metode yang efektif untuk diagnosis diabetes.	yang dapat mempengaruhi performa prediksi.	
5	Klasifikasi Diabetes Menggunakan Algoritma Support Vector Machine Radial Basis Function	Muhammad Hilmy Haidar Aly. Teknologi Informasi dan Multimedia pada tahun 2023	Menganalisis dan mengimplementasikan algoritma Support Vector Machine dengan kernel Radial Basis Function untuk klasifikasi diabetes, serta membandingkan kinerjanya dengan algoritma lain seperti K-Nearest Neighbor	Penggunaan algoritma SVM dengan kernel RBF dalam klasifikasi diabetes mampu mencapai akurasi maksimal sebesar 87%. Hal ini menunjukkan potensi algoritma SVM-RBF dalam mendeteksi diabetes secara efektif.	Hyperparameter tuning untuk parameter *C* dan *gamma* tidak diterapkan, sehingga optimasi model kurang maksimal. Selain itu, preprocessing data hanya menggunakan *Min-Max Scaler* tanpa teknik tambahan, yang dapat membatasi kualitas normalisasi fitur.	Penelitian ini mengoptimasi parameter C, gamma, dan jumlah fitur menggunakan PSO. Hasil menunjukkan model lebih stabil dan sensitif dalam mendeteksi diabetes.
6	Klasifikasi Menggunakan Metode Support Vector Machine dan Random Forest untuk Deteksi Awal Risiko Diabetes Melitus	Chea Zahrah Vaganza Junus, Tarno Tarno, dan Puspita Kartikasari. Teknologi Informasi, Komunikasi dan Industri pada tahun 2022	Membandingkan performa algoritma Support Vector Machine dan Random Forest dalam mendeteksi awal risiko diabetes melitus berdasarkan dataset yang digunakan dan ingin mengetahui algoritma mana yang lebih efektif untuk prediksi dini diabetes.	Penelitian menunjukkan bahwa kernel non-linear Dalam SVM, khususnya kombinasi RBF dan RBF City Block, secara signifikan meningkatkan akurasi prediksi diabetes tipe 2 dibandingkan metode konvensional. Hasilnya mencatat akurasi 85.5%, dengan recall 87.0%, precision 83.4%,	Dataset yang digunakan terdiri dari 520 data, namun tidak dilakukan preprocessing yang jelas, sehingga kualitas data kurang optimal. Selain itu, *SVM* tidak menerapkan *hyperparameter tuning*, yang dapat mempengaruhi akurasi model.	Penelitian ini menerapkan pipeline preprocessing terintegrasi dan optimasi threshold. Hasil menunjukkan keseimbangan yang lebih baik antara sensitivitas dan ketepatan prediksi.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
				F1-score 85.2%, dan AUC 85.5%. Hal ini menegaskan potensi metode ini dalam aplikasi klinis untuk deteksi dini dan pengelolaan risiko diabetes, yang dapat membantu mengurangi dampak penyakit ini secara efektif.		

2.3. Landasan Teori

Dasar teori adalah pijakan ilmiah yang digunakan untuk mendukung penelitian ini, sekaligus memberikan pemahaman yang mendalam tentang konsep, metode, dan teknologi terkait. Penelitian ini berpusat pada pemodelan prediksi Diabetes Mellitus menggunakan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF). Oleh karena itu, berikut ini adalah dasar teori yang disajikan oleh peneliti :

2.3.1. Diabetes Militus

Diabetes mellitus merupakan salah satu masalah kesehatan utama di dunia yang ditandai dengan tingginya kadar gula darah akibat gangguan produksi atau fungsi insulin. Penyakit ini menjadi perhatian serius karena dapat menyebabkan berbagai komplikasi jangka panjang, seperti kerusakan saraf, penyakit kardiovaskular, gagal ginjal, hingga kebutaan, yang dapat membahayakan kehidupan penderitanya. Menurut data International Diabetes Federation tahun 2019, Indonesia berada di peringkat kedua di wilayah Pasifik Barat dengan 10,7 juta penderita diabetes, atau sekitar 6,2% dari populasi dewasa. Tingginya angka tersebut mendorong perlunya deteksi dini sebagai langkah penting untuk mengendalikan peningkatan prevalensi diabetes.

2.3.2. Missing Value

Dijelaskan oleh Kuhn & Johnson (Kuhn & Johnson, 2013) *missing value* adalah keadaan ketika suatu nilai tidak tersedia atau hilang untuk sebuah variabel

dalam dataset. Hal ini sering kali disebabkan oleh kesalahan dalam proses pengumpulan, penyimpanan, atau entri data, serta situasi di mana responden memilih untuk tidak memberikan informasi. Kehadiran *missing value* dapat berdampak buruk pada analisis data dan algoritma pemodelan dengan mengurangi *accuracy* dan meningkatkan risiko bias jika tidak ditangani dengan tepat (Han & Kamber, 2012).

Untuk menangani *missing value*, tersedia berbagai teknik. Metode sederhana meliputi penghapusan baris yang memiliki *missing value* atau mengganti nilai yang hilang dengan nilai rata-rata atau median. Metode yang lebih kompleks menggunakan model regresi, inferensi Bayesian, atau *decision tree* untuk memperkirakan nilai yang paling mungkin berdasarkan data yang ada. Meskipun metode ini dapat meningkatkan kualitas dataset, pendekatan ini juga dapat menjadikan dataset menjadi bias, karena nilai yang diestimasi tidak selalu mencerminkan data sebenarnya. Oleh karena itu, penting untuk menyeimbangkan kelayakan saat memilih strategi imputasi (Han & Kamber, 2012).

Penanganan *missing value* sangat penting untuk memastikan analisis data yang akurat dan pengambilan keputusan yang andal. Dilengkapi oleh Kuhn (Kuhn & Johnson, 2013) jika *missing value* diabaikan atau tidak ditangani dengan baik, hal ini dapat menyebabkan hasil yang bias, keandalan model yang berkurang, dan kesimpulan yang keliru. Dengan menangani *missing value*, konsistensi dan kelengkapan data dapat ditingkatkan, sehingga model *Machine Learning* menjadi lebih efektif. Oleh karena itu, langkah penanganan *missing value* merupakan bagian krusial dari proses *preprocessing* data, untuk menjaga integritas dan kegunaan data

2.3.3. *Split Dataset*

Dalam proses pembangunan model machine learning, *split* dataset sangat penting. Dijabarkan oleh Geron (Géron, 2019) *split* dataset merupakan proses membagi data menjadi dua atau lebih, umumnya berupa data *training* dan data *testing*. Data *training* digunakan untuk melatih model *Machine Learning*, sedangkan data *testing* digunakan untuk mengevaluasi performa model yang telah dilatih. Pemisahan dataset bertujuan untuk memastikan bahwa model dapat memprediksi data baru dengan baik, bukan hanya menghafal data *training*. Proporsi umum yang digunakan adalah 80% data *training* dan 20% data *testing*, meskipun proporsi ini dapat bervariasi tergantung pada ukuran dataset dan tujuan analisis. Di mana bagian terbesar digunakan untuk training guna memaksimalkan proses pembelajaran model. Dalam beberapa kasus, data *validation* juga digunakan sebagai subset tambahan untuk mengoptimalkan *hyperparameter* tanpa melibatkan data *testing*. Teknik ini mendukung pendekatan yang disebut *train-test split validation*, yang merupakan langkah awal penting untuk membangun model yang dapat digeneralisasi. Pembagian dataset memastikan bahwa model tidak mengalami masalah *overfitting* atau *underfitting* pada data.

Pentingnya pembagian dataset terletak pada kemampuan model untuk menggeneralisasi ke data baru. Tanpa proses ini, ada risiko model menjadi *overfit* terhadap data *training* dan gagal memprediksi secara akurat pada data yang belum pernah dilihat sebelumnya (Kuhn & Johnson, 2013). Selain itu, pada saat *training* memberikan peluang untuk menyesuaikan parameter model secara optimal,

sedangkan saat *testing* memberikan penilaian performa akhir pada data yang benar-benar tidak terlihat selama pelatihan

2.3.4. Balancing Dataset dengan Metode Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (Synthetic Minority Over-sampling Technique) merupakan sebuah pendekatan pembelajaran yang dikembangkan untuk mengatasi permasalahan ketidakseimbangan kelas dalam dataset, sebagaimana diperkenalkan oleh (Chawla et al., 2002). Ketidakseimbangan kelas terjadi ketika jumlah data pada satu kelas jauh lebih sedikit dibandingkan kelas lainnya, sehingga berpotensi menimbulkan bias pada proses pelatihan model Machine Learning. Untuk mengatasi hal tersebut, SMOTE menghasilkan sampel sintetis dari kelas minoritas dengan memanfaatkan data yang sudah ada.

Berbeda dengan metode oversampling konvensional yang hanya menggandakan data minoritas, SMOTE menciptakan data baru melalui proses interpolasi. Metode ini bekerja dengan memilih sampel minoritas secara acak, kemudian membentuk titik baru di antara pasangan sampel minoritas berdasarkan perhitungan jarak Euclidean pada ruang fitur. Pendekatan ini memungkinkan distribusi kelas minoritas menjadi lebih merata dan informatif dalam proses klasifikasi.

Inti dari mekanisme SMOTE terletak pada interpolasi linear, di mana nilai fitur dari data sintetis dihasilkan di antara dua sampel minoritas yang berdekatan. Dengan cara ini, representasi kelas minoritas dapat diperluas tanpa sekadar

menambah duplikasi data yang berpotensi menyebabkan overfitting. Penerapan SMOTE terbukti mampu meningkatkan performa model pada berbagai permasalahan klasifikasi yang sensitif terhadap ketidakseimbangan kelas, seperti deteksi penipuan, prediksi medis, dan diagnosis penyakit.

Meskipun demikian, SMOTE memiliki keterbatasan, terutama ketika data minoritas mengandung noise. Dalam kondisi tersebut, data sintetis yang dihasilkan berisiko kurang relevan atau tidak merepresentasikan distribusi sebenarnya. Oleh karena itu, SMOTE sering dikombinasikan dengan teknik tambahan, seperti strategi pembersihan data (cleaning strategies) atau metode ensemble, untuk meningkatkan efektivitas dan keandalan model.

Berbagai penelitian menunjukkan keunggulan SMOTE dalam aplikasi nyata. (He, 2013b) menjelaskan penerapan SMOTE dalam beragam bidang, mulai dari pengenalan pola hingga klasifikasi dokumen. Sementara itu, (Zaki & Meira, Jr, 2020) menekankan bahwa karena SMOTE membangkitkan data sintetis berdasarkan pola yang ada, penggunaannya memerlukan pemahaman mendalam terhadap karakteristik dataset agar risiko overfitting dapat diminimalkan dan data yang dihasilkan tetap representatif terhadap distribusi kelas minoritas yang sebenarnya.

2.3.5. *Particle Swarm Optimization (PSO)*

Particle Swarm Optimization (PSO) merupakan algoritma optimasi berbasis populasi yang terinspirasi dari perilaku sosial kawanan, seperti burung atau ikan, dalam mencari sumber makanan. Dalam PSO, setiap solusi direpresentasikan

sebagai partikel yang bergerak di dalam ruang pencarian dengan kecepatan tertentu. Pergerakan partikel dipengaruhi oleh dua informasi utama, yaitu pengalaman terbaik partikel itu sendiri (personal best/pbest) dan pengalaman terbaik seluruh populasi (global best/gbest). Melalui mekanisme ini, partikel secara kolektif mengeksplorasi ruang solusi dan secara bertahap bergerak menuju solusi optimal.(Kennedy & Eberhart, 2022)

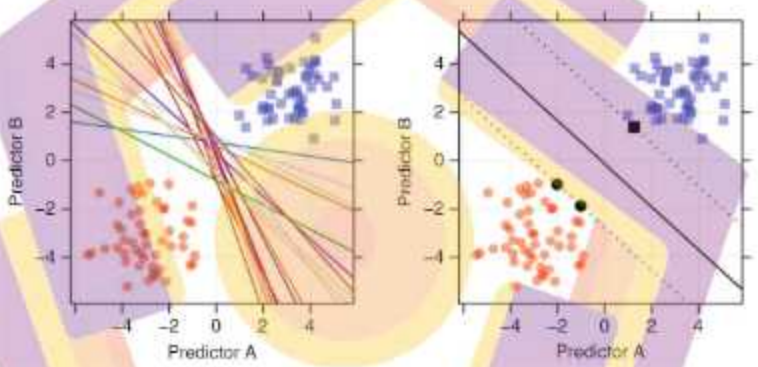
Keunggulan PSO terletak pada strukturnya yang sederhana, jumlah parameter yang relatif sedikit, serta kemampuan konvergensi yang cepat. PSO tidak memerlukan operasi kompleks seperti crossover dan mutasi sebagaimana pada algoritma genetika, sehingga lebih efisien secara komputasi. Dalam penelitian ini, PSO digunakan untuk mengoptimasi parameter model Support Vector Machine serta jumlah fitur terbaik, sehingga diperoleh konfigurasi model yang optimal dan memiliki kemampuan generalisasi yang baik.

2.3.6. *Support Vector Machine (SVM) dengan Kernel Radial Basis Function* (RBF)

Menurut (Kuhn & Johnson, 2013), Support Vector Machines (SVM) adalah salah satu model statistika yang pertama kali diperkenalkan pada pertengahan 1960-an oleh Vladimir Vapnik. SVM dirancang untuk menemukan hyperplane yang memisahkan data dengan margin maksimum, di mana hyperplane ini berfungsi sebagai batas keputusan yang membedakan antara kelas-kelas dalam dataset.

Sebagai ilustrasi, dalam contoh yang dijelaskan (Kuhn & Johnson, 2013), ditunjukkan bahwa dua variabel digunakan untuk memprediksi dua kelas yang

dapat sepenuhnya dipisahkan. Dalam situasi tersebut, terdapat banyak kemungkinan garis pemisah linier yang dapat memisahkan data dengan sempurna. Namun, tantangannya adalah menentukan garis pemisah yang paling optimal. Dalam kasus ini, akurasi mungkin tidak menjadi metrik yang memadai, karena semua garis pemisah akan dianggap memiliki kinerja yang sama baiknya. Oleh karena itu, diperlukan metrik lain yang lebih sesuai untuk mengevaluasi sejauh mana sebuah model dapat bekerja secara efektif.



Gambar 2.1. Data dengan garis pemisah

Vladimir Vapnik menawarkan solusi dengan memperkenalkan konsep margin, yaitu jarak antara garis pemisah (batas klasifikasi) dengan titik data terdekat. Margin ini terlihat jelas pada ilustrasi dengan garis solid sebagai garis pemisah dan garis putus-putus yang menunjukkan jarak maksimum ke data terdekat. Dalam contoh tersebut, terdapat tiga titik data berjarak sama dari garis pemisah, yang dikenal sebagai margin. SVM memanfaatkan konsep ini untuk mencari hyperplane yang memaksimalkan margin, menjadikannya batas keputusan yang efektif dalam membedakan kelas-kelas data. Titik-titik data yang paling dekat

dengan margin disebut support vectors, yang berperan penting dalam menentukan posisi hyperplane.

Penelitian (Sharma et al., 2023) mengidentifikasi bahwa banyak model prediksi sebelumnya berfokus pada diabetes gestasional yang hanya mencakup pasien wanita, sehingga akurasinya terbatas. Dalam studi ini, peneliti memperluas dataset dengan informasi dari pasien pria dan wanita serta mengimprovisasi model SVM agar lebih akurat dalam mendeteksi risiko diabetes berdasarkan parameter medis yang lebih luas. Metodologi yang digunakan melibatkan pengumpulan data dari UCI Machine Learning Repository, pemrosesan data termasuk konversi nilai string ke biner, dan penerapan model SVM untuk klasifikasi. Hasil penelitian menunjukkan bahwa model yang diusulkan mencapai akurasi 93,26%, lebih tinggi dibandingkan metode sebelumnya. Studi ini menekankan pentingnya penggunaan teknik *machine learning* dalam deteksi dini penyakit kronis, yang memungkinkan pasien mendapatkan diagnosis lebih cepat dan tepat waktu tanpa harus mengunjungi fasilitas medis secara langsung.

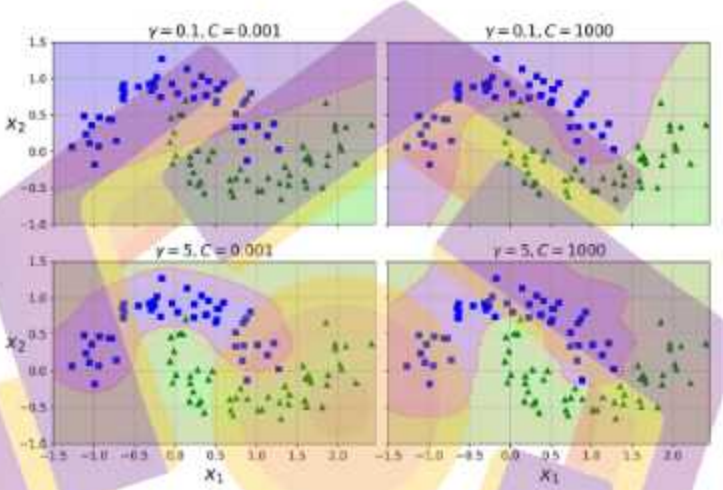
Meskipun berbagai model prediksi diabetes telah dikembangkan, sebagian besar masih menggunakan dataset terbatas, seperti *Pima Indian Diabetes Dataset*, dan kurang memperhatikan interpretabilitas keputusan model. Untuk mengatasi keterbatasan ini, penelitian ini menggunakan dataset gabungan, termasuk dataset pribadi yang diperoleh dari pekerja industri tekstil di Bangladesh. Peneliti menerapkan algoritma *mutual information* untuk pemilihan fitur serta metode *semi-supervised learning* menggunakan *Extreme Gradient Boosting* (XGBoost) untuk menangani data yang tidak lengkap. Selain itu, untuk mengatasi ketidakseimbangan

data, digunakan teknik *SMOTE* dan *ADASYN*. Berbagai algoritma ML, seperti *Decision Tree*, *SVM*, *Random Forest*, *Logistic Regression*, dan teknik ensemble diterapkan untuk mencari model terbaik. Hasil penelitian menunjukkan bahwa kombinasi XGBoost dengan *ADASYN* memberikan akurasi terbaik sebesar 81% dengan skor AUC 0.84. Untuk meningkatkan transparansi prediksi, diterapkan metode XAI menggunakan *LIME* dan *SHAP* yang memungkinkan pengguna memahami faktor-faktor yang memengaruhi keputusan model. Akhirnya, model terbaik ini diintegrasikan ke dalam aplikasi web dan Android untuk memungkinkan prediksi diabetes secara real-time. (Tasin et al., 2023)

Ketika data tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel, seperti Radial Basis Function (RBF), polinomial, atau sigmoid, untuk mentransformasikan data ke ruang berdimensi lebih tinggi agar dapat dipisahkan dengan lebih mudah. Kernel trick ini memungkinkan SVM bekerja secara efisien tanpa memerlukan transformasi langsung yang membutuhkan banyak komputasi. Kernel RBF, misalnya, sering digunakan karena kemampuannya menangkap pola-pola non-linear dan performa yang baik pada data berdimensi tinggi.

Dalam penerapan kernel RBF, dua parameter penting yang harus dioptimalkan adalah C dan gamma (γ). Parameter C menentukan toleransi model terhadap kesalahan klasifikasi, di mana nilai tinggi cenderung menghasilkan model yang lebih ketat dan rentan overfitting, sementara nilai rendah memberikan margin yang lebih besar tetapi meningkatkan risiko underfitting. Parameter gamma mengontrol sejauh mana pengaruh dari setiap titik data, di mana nilai tinggi membuat model lebih fokus pada data terdekat, menghasilkan keputusan yang

kompleks, sedangkan nilai rendah menghasilkan model yang lebih sederhana dan stabil. Dengan penyesuaian yang tepat, SVM dengan kernel RBF mampu menangkap pola data yang kompleks secara efektif, memberikan performa yang unggul dalam klasifikasi data non-linear.



Gambar 2.2. SVM dengan Kernel RBF

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tujuan membangun dan mengevaluasi model klasifikasi untuk memprediksi penyakit diabetes. Metode utama yang digunakan adalah Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) yang dioptimasi menggunakan Particle Swarm Optimization (PSO).

Pendekatan ini dipilih karena SVM memiliki kemampuan tinggi dalam memodelkan hubungan nonlinier antar variabel medis, sedangkan PSO berfungsi sebagai algoritma optimasi global yang mampu mencari kombinasi parameter terbaik secara adaptif. Kombinasi kedua metode ini diharapkan mampu menghasilkan model prediksi diabetes yang lebih akurat dan stabil.

Seluruh proses pemodelan dibangun menggunakan pendekatan pipeline leakage-free, di mana seluruh tahapan preprocessing, penyeimbangan data, seleksi fitur, dan pelatihan model hanya dilakukan pada data training. Dengan pendekatan ini, data testing benar-benar tidak terlibat dalam proses pembelajaran maupun optimasi, sehingga evaluasi akhir mencerminkan performa model yang sesungguhnya.

3.2. Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset, yang terdiri dari data medis pasien perempuan dengan delapan atribut klinis dan satu label target. Dataset ini banyak digunakan dalam penelitian prediksi diabetes karena memuat indikator medis yang relevan dan telah tervalidasi secara luas.

Tabel 3.2 Delapan atribut yang digunakan sebagai variabel input adalah:

No	Nama Atribut	Deskripsi Singkat
1	**Pregnancies**	Jumlah kehamilan yang pernah dialami pasien
2	**Glucose**	Kadar glukosa plasma (tes toleransi glukosa)
3	**BloodPressure**	Tekanan darah diastolik (mm Hg)
4	**SkinThickness**	Ketebalan lipatan kulit triceps (mm)
5	**Insulin**	Kadar insulin serum (mu U/ml)
6	**BMI**	Indeks massa tubuh (kg/m²)
7	**DiabetesPedigreeFunction**	Indikator riwayat diabetes dalam keluarga
8	**Age**	Usia pasien (tahun)

Label target adalah Outcome, dengan nilai 1 menunjukkan pasien menderita diabetes dan 0 menunjukkan pasien tidak menderita diabetes. Dataset ini memiliki distribusi kelas yang tidak seimbang, sehingga diperlukan teknik khusus untuk menghindari bias model terhadap kelas mayoritas.model.

3.3. Pembagian Dataset (Split Data)

Pembagian dataset dalam penelitian ini dilakukan menggunakan metode stratified train–test split, di mana data dibagi menjadi dua bagian utama, yaitu 75% sebagai data training dan 25% sebagai data testing. Pendekatan stratifikasi digunakan agar distribusi kelas pada kedua subset tetap mencerminkan distribusi kelas pada dataset asli. Dengan cara ini, proporsi pasien diabetes dan non-diabetes pada data training dan data testing tetap seimbang, sehingga model tidak dilatih atau diuji pada distribusi yang bias.

Data training berperan sebagai dasar utama dalam seluruh proses pembelajaran model. Pada bagian ini dilakukan seluruh tahapan pemodelan, mulai dari preprocessing, penyeimbangan data, seleksi fitur, hingga proses optimasi parameter menggunakan Particle Swarm Optimization (PSO). Dengan hanya menggunakan data training dalam tahap-tahap tersebut, model dapat belajar pola hubungan antara fitur medis dan status diabetes secara optimal tanpa terpengaruh oleh data yang seharusnya digunakan untuk pengujian.

Sementara itu, data testing disimpan secara terpisah dan tidak pernah digunakan dalam proses pelatihan maupun optimasi. Data ini hanya digunakan pada tahap evaluasi akhir untuk mengukur kinerja model secara objektif. Pendekatan ini bertujuan untuk mencegah terjadinya data leakage, yaitu kondisi di mana informasi dari data uji secara tidak langsung mempengaruhi proses pelatihan, yang dapat menyebabkan hasil evaluasi menjadi terlalu optimistis dan tidak mencerminkan performa model pada data baru.

3.4. Arsitektur Pipeline Pemodelan

Seluruh proses pemodelan dalam penelitian ini dirancang menggunakan sebuah pipeline terintegrasi yang menggabungkan seluruh tahapan pemrosesan data dan pembelajaran model dalam satu alur kerja yang konsisten. Pipeline ini terdiri dari empat komponen utama, yaitu

StandardScaler untuk normalisasi data,

SMOTE untuk penyeimbangan kelas,

SelectKBest untuk seleksi fitur, dan

Support Vector Machine (SVM) sebagai algoritma klasifikasi.

Penggunaan pipeline memastikan bahwa setiap proses preprocessing diterapkan secara identik pada setiap fold dalam cross-validation maupun pada setiap evaluasi partikel dalam Particle Swarm Optimization (PSO). Artinya, ketika model dievaluasi, data terlebih dahulu distandarisasi, kemudian diseimbangkan, lalu dipilih fitur terbaiknya, sebelum akhirnya digunakan untuk melatih model SVM. Pendekatan ini mencegah terjadinya inkonsistensi atau kesalahan dalam urutan pemrosesan data yang dapat memengaruhi hasil pelatihan.

Selain itu, pipeline juga berperan penting dalam mencegah terjadinya data leakage, karena semua tahapan preprocessing hanya diterapkan pada data training di dalam proses validasi silang dan optimasi parameter. Data testing tidak pernah terlibat dalam pembentukan skala, pembangkitan data sintetis, maupun pemilihan fitur. Dengan demikian, evaluasi kinerja model yang diperoleh benar-benar mencerminkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. training.

3.5. Tahapan Preprocessing Data

Tahapan preprocessing data merupakan langkah penting dalam penelitian ini untuk memastikan bahwa data yang digunakan berada dalam kondisi optimal sebelum dilakukan proses pelatihan model klasifikasi. Dataset yang digunakan memiliki karakteristik numerik dengan perbedaan skala antar fitur serta distribusi kelas yang tidak seimbang, sehingga diperlukan serangkaian proses praproses agar model Support Vector Machine (SVM) dapat bekerja secara efektif dan menghasilkan kinerja yang optimal.

3.5.1. Standardisasi Data dengan StandardScaler

Tahap pertama dalam pipeline adalah StandardScaler, yang berfungsi untuk menormalkan seluruh fitur agar memiliki rata-rata nol dan standar deviasi satu. Standardisasi diperlukan karena algoritma SVM dengan kernel RBF sangat sensitif terhadap perbedaan skala antar fitur. Tanpa normalisasi, fitur dengan rentang nilai besar dapat mendominasi perhitungan jarak dalam ruang fitur dan menyebabkan model menjadi bias.

3.5.2. Penyeimbangan Data dengan SMOTE

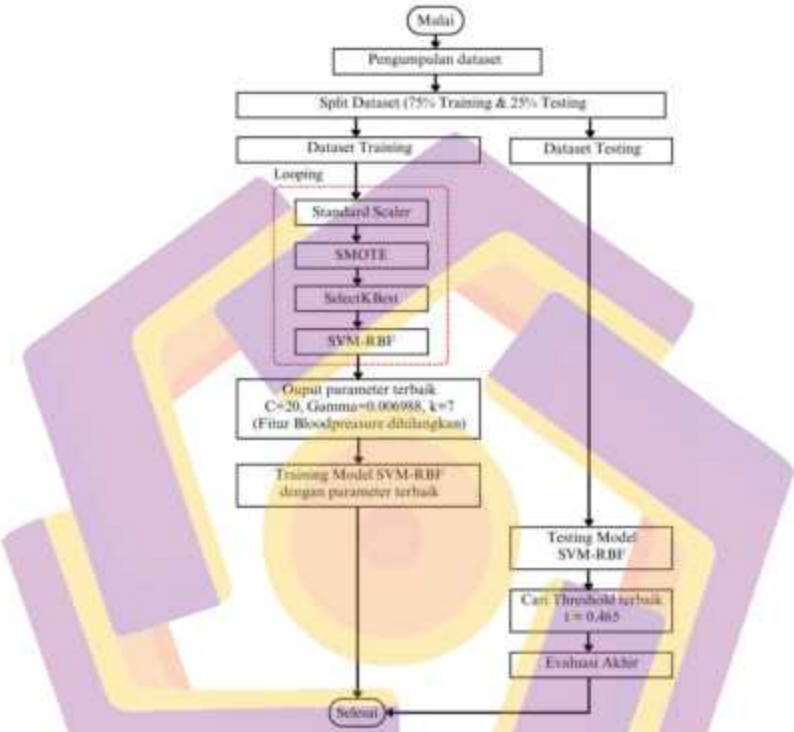
Setelah seluruh fitur dinormalisasi menggunakan StandardScaler, tahap berikutnya dalam pipeline adalah penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Teknik ini dirancang untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset, di mana jumlah data pasien non-diabetes lebih banyak dibandingkan pasien diabetes. SMOTE

bekerja dengan membentuk sampel sintetis pada kelas minoritas berdasarkan kedekatan antar titik data dalam ruang fitur. Dengan cara ini, data baru yang dihasilkan tidak sekadar hasil duplikasi, tetapi merupakan interpolasi dari sampel yang ada, sehingga dapat memperkaya representasi pola pada kelas minoritas.

SMOTE diterapkan sebagai bagian dari pipeline yang hanya bekerja pada data training dan di dalam proses cross-validation serta optimasi PSO. Artinya, data testing sama sekali tidak terlibat dalam proses pembentukan data sintetis. Pendekatan ini sangat penting untuk mencegah terjadinya data leakage, yaitu kondisi di mana informasi dari data uji secara tidak langsung mempengaruhi proses pelatihan model. Dengan menempatkan SMOTE di dalam pipeline, setiap fold pada cross-validation melakukan oversampling hanya pada subset training-nya masing-masing, sehingga evaluasi performa model tetap objektif dan mencerminkan kemampuan generalisasi model pada data baru.

Penerapan SMOTE bertujuan untuk mengurangi bias model terhadap kelas mayoritas dengan menyeimbangkan distribusi data pada kelas minoritas. Tanpa penyeimbangan data, model klasifikasi berpotensi menghasilkan keputusan yang kurang representatif terhadap kasus diabetes. Oleh karena itu, SMOTE dikombinasikan dengan standardisasi fitur, seleksi fitur, dan optimasi hyperparameter menggunakan PSO untuk membangun model yang lebih stabil dan memiliki kemampuan generalisasi yang baik.

3.5.3. Seleksi Fitur dengan SelectKBest



Gambar 3.1 Alur Penelitian

Berdasarkan pada gambar di atas bahwa alur penelitian disusun secara sistematis dimulai dari input dataset *Pima Indians Diabetes* yang kemudian dipartisi menjadi data latih (75%) dan data uji (25%) melalui teknik *stratified sampling*. Proses inti dari sistem ini berfokus pada tahap optimasi model menggunakan algoritma *Particle Swarm Optimization* (PSO), yang bertugas mencari solusi terbaik secara simultan terhadap tiga variabel kunci: parameter regularisasi SVM (C), koefisien kernel RBF (gamma), dan jumlah fitur optimal (k). Dalam setiap iterasinya, algoritma membangun sebuah *pipeline* dinamis yang

mencakup standardisasi data, penanganan ketidakseimbangan kelas (SMOTE), serta seleksi fitur *SelectKBest*, yang kinerjanya divalidasi menggunakan *5-Fold Stratified Cross-Validation*. Pendekatan terintegrasi ini memungkinkan sistem untuk mengeliminasi fitur *noise* secara otomatis (menghasilkan k=7) sekaligus menemukan konfigurasi *hyperparameter* yang paling presisi.

Setelah parameter global terbaik (Global Best) ditemukan, model SVM dilatih kembali pada keseluruhan data latih untuk menangkap pola data secara utuh. Model final ini kemudian digunakan untuk memprediksi probabilitas pada data uji yang disimpan (*hold-out*), di mana hasilnya diproses lebih lanjut melalui tahap optimasi *threshold*. Pada tahap ini, ambang batas keputusan (*decision boundary*) digeser secara iteratif untuk menemukan nilai *threshold* spesifik (t=0.465) yang mampu menyeimbangkan akurasi dan sensitivitas. Rangkaian proses ini diakhiri dengan evaluasi performa final menggunakan metrik Akurasi, *Recall*, dan AUC, guna memastikan keandalan model dalam mendeteksi diabetes.

Algoritma Particle Swarm Optimization (PSO) tidak hanya bertugas mencari parameter internal model SVM (C dan gamma), tetapi juga melakukan seleksi fitur secara otomatis. Dalam desain ini, setiap partikel dalam populasi PSO merepresentasikan sebuah vektor solusi tiga dimensi (C, gamma, k).

Variabel k dalam struktur partikel ini didefinisikan sebagai parameter input untuk fungsi SelectKBest, yang menentukan jumlah fitur paling informatif yang akan dipertahankan berdasarkan skor statistik ANOVA. Penting untuk dibedakan bahwa variabel k dalam ruang pencarian optimasi ini bukanlah representasi dari jumlah lipatan (fold) dalam validasi silang. Proses validasi model tetap menggunakan standar baku 5-Fold Stratified Cross-Validation dengan nilai konstan yang tidak diubah selama proses optimasi berjalan.

Adapun struktur representasi partikel yang diusulkan adalah sebagai berikut:

1. Dimensi 1 (C): Parameter regularisasi untuk mengontrol margin error SVM (Tipe: Kontinu).
2. Dimensi 2 (gamma): Koefisien kernel RBF yang mengatur cakupan pengaruh data latih (Tipe: Kontinu).
3. Dimensi 3 (k): Jumlah fitur terpilih dari total 8 atribut dataset Pima (Tipe: Diskrit/Integer, rentang 5 sampai 8).

Dengan memasukkan k ke dalam fungsi fitness, algoritma dipaksa untuk mencari keseimbangan terbaik: model harus memiliki akurasi tinggi namun dengan jumlah fitur yang efisien. Jika nilai k yang ditemukan lebih kecil dari total fitur (misal k < 8), hal ini mengindikasikan bahwa algoritma berhasil mengeliminasi fitur yang bersifat noise atau redundan yang justru dapat menurunkan performa klasifikasi.

3.6. Optimasi Parameter SVM Menggunakan Particle Swarm Optimization (PSO)

Pada tahap ini, optimasi parameter Support Vector Machine (SVM) dilakukan menggunakan Particle Swarm Optimization (PSO) untuk memperoleh kombinasi parameter yang mampu menghasilkan kinerja klasifikasi yang optimal. Dalam skema ini, setiap partikel pada PSO merepresentasikan satu kandidat solusi yang terdiri dari tiga parameter utama, yaitu parameter regularisasi C, parameter kernel gamma pada fungsi RBF, serta jumlah fitur k yang dipilih oleh metode SelectKBest. Dengan pendekatan ini, PSO tidak hanya berfungsi untuk menyesuaikan kompleksitas dan fleksibilitas model SVM, tetapi juga secara

simultan mengoptimalkan dimensi fitur yang digunakan, sehingga model dapat dibangun secara lebih efisien dan informatif.

Nilai awal setiap partikel diinisialisasi secara acak dalam batas tertentu untuk masing-masing parameter, kemudian setiap konfigurasi dievaluasi menggunakan fungsi fitness yang menggabungkan nilai akurasi dan ROC-AUC yang diperoleh dari 5-fold Stratified Cross Validation pada data training. Penggunaan validasi silang yang terstratifikasi memastikan bahwa distribusi kelas tetap seimbang pada setiap fold, sehingga penilaian kinerja model menjadi lebih representatif dan stabil. Kombinasi dua metrik tersebut dalam fungsi fitness bertujuan untuk menyeimbangkan antara ketepatan prediksi dan kemampuan model dalam membedakan kelas positif dan negatif.

Selama proses iteratif PSO, setiap partikel memperbarui posisinya berdasarkan dua komponen utama, yaitu pengalaman terbaik yang pernah dicapainya sendiri dan pengalaman terbaik yang dicapai oleh seluruh populasi partikel. Mekanisme ini memungkinkan partikel bergerak menuju wilayah ruang solusi yang memberikan nilai fitness lebih tinggi, sekaligus tetap menjaga eksplorasi agar tidak terjebak pada solusi lokal.

3.7. Pelatihan Model

Setelah parameter optimal diperoleh melalui proses optimasi PSO, model Support Vector Machine (SVM) dibangun kembali menggunakan pipeline yang sama seperti pada tahap optimasi, yang mencakup standardisasi data, penyeimbangan kelas dengan SMOTE, seleksi fitur, dan pembelajaran model.

Penggunaan pipeline yang konsisten memastikan bahwa seluruh tahapan pemrosesan data diterapkan dengan urutan dan konfigurasi yang identik, sehingga tidak terjadi perbedaan perlakuan antara proses optimasi dan pelatihan akhir. Seluruh data training kemudian digunakan untuk melatih model dengan konfigurasi terbaik tersebut agar informasi yang tersedia dapat dimanfaatkan secara maksimal.

Tahap pelatihan ini bertujuan untuk membentuk model klasifikasi yang mampu menangkap pola hubungan antara atribut klinis pasien, seperti kadar glukosa, indeks massa tubuh, dan usia, dengan status diabetes. Dengan memanfaatkan parameter dan fitur yang telah dioptimasi secara global, model diharapkan memiliki kemampuan generalisasi yang baik ketika diterapkan pada data baru, sehingga dapat memberikan prediksi yang lebih andal dalam membedakan pasien yang menderita diabetes dan yang tidak.

3.8 Optimasi Threshold Klasifikasi

Pada tahap ini, hasil keluaran model SVM tidak langsung digunakan dalam bentuk kelas biner, melainkan berupa nilai probabilitas yang menunjukkan tingkat keyakinan model terhadap suatu sampel termasuk ke dalam kelas diabetes. Oleh karena itu, dilakukan proses optimasi nilai threshold untuk menentukan batas probabilitas yang paling tepat dalam mengklasifikasikan pasien sebagai diabetes atau non-diabetes. Dengan memvariasikan nilai threshold dalam suatu rentang tertentu, model dapat dievaluasi pada berbagai titik keputusan yang berbeda, sehingga tidak bergantung hanya pada batas standar 0,5.

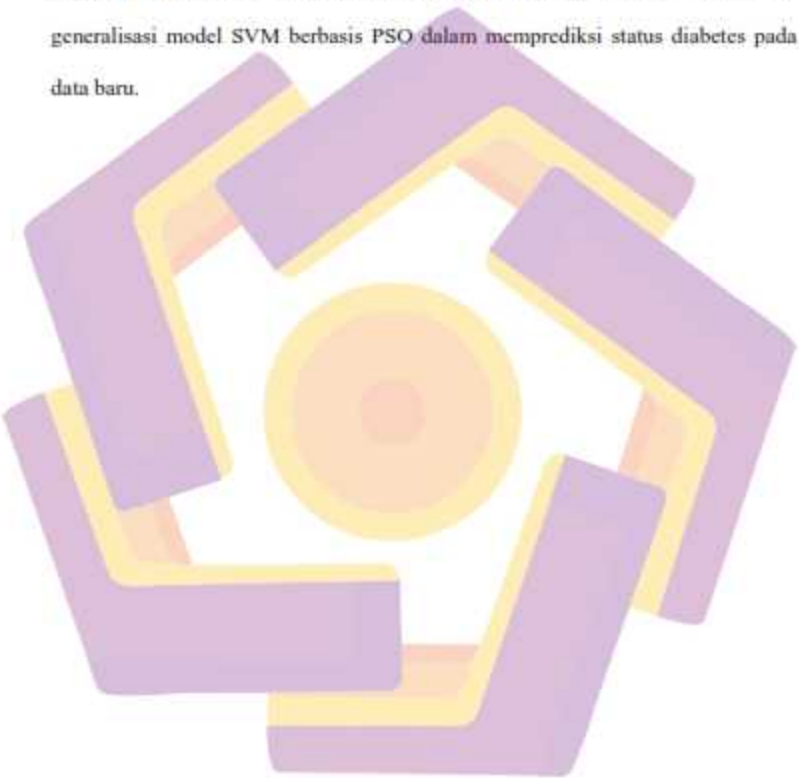
Pendekatan ini penting karena dataset yang digunakan memiliki distribusi kelas yang tidak seimbang. Dengan mengatur threshold secara optimal, diharapkan model mampu mencapai kesimbangan yang lebih baik antara kemampuan mendeteksi pasien yang benar-benar menderita diabetes (recall) dan ketepatan dalam memprediksi kasus diabetes (precision). Tahap ini membantu memastikan bahwa model tidak hanya mengejar akurasi semata, tetapi juga memiliki kinerja yang lebih adil dan relevan secara klinis dalam konteks deteksi penyakit.

3.9 Testing Model

Pada tahap testing, model yang digunakan adalah model SVM hasil training terakhir yang telah dibangun menggunakan parameter terbaik hasil optimasi PSO serta pipeline preprocessing yang sama. Model ini sebelumnya telah dilatih menggunakan seluruh data training melalui pipeline yang mencakup standardisasi, penyeimbangan data dengan SMOTE, seleksi fitur, dan pembelajaran SVM. Dengan demikian, model yang diuji bukan model sementara dari proses cross-validation, melainkan model final yang telah teroptimasi secara global.

Data testing kemudian dimasukkan ke dalam model tersebut tanpa melalui proses pelatihan ulang. Pipeline yang sama secara otomatis menerapkan standardisasi dan seleksi fitur berdasarkan parameter yang telah ditentukan sebelumnya, tetapi tanpa melakukan SMOTE, karena data testing harus tetap merepresentasikan distribusi asli. Model kemudian menghasilkan probabilitas prediksi untuk setiap sampel, yang selanjutnya dikonversi menjadi label kelas berdasarkan threshold terbaik yang telah ditentukan pada tahap optimasi threshold.

Hasil prediksi pada data testing inilah yang digunakan untuk menghitung metrik evaluasi seperti accuracy, precision, recall, F1-score, dan ROC-AUC. Karena data testing tidak terlibat dalam proses optimasi PSO, pemilihan fitur, maupun pelatihan model, maka hasil evaluasi ini mencerminkan kemampuan generalisasi model SVM berbasis PSO dalam memprediksi status diabetes pada data baru.



BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Pada penelitian ini, dilakukan pemodelan prediksi diabetes menggunakan algoritma **Support Vector Machine (SVM)** dengan kernel **Radial Basis Function (RBF)**. Tahapan penelitian dimulai dari pengumpulan data, preprocessing, pelatihan model, hingga evaluasi kinerja model. Berikut hasil dan pembahasannya.

4.1 Karakteristik Dataset

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset yang terdiri dari 768 data pasien perempuan dengan delapan atribut klinis dan satu label target, yaitu *Outcome*. Dataset ini banyak digunakan dalam penelitian klasifikasi medis karena seluruh fitur mewakili indikator biologis yang berkaitan langsung dengan risiko diabetes mellitus. Seluruh atribut bersifat numerik, sehingga sangat sesuai untuk diproses menggunakan algoritma berbasis jarak seperti Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF).

Berdasarkan hasil eksplorasi awal menggunakan fungsi info() dan value_counts(), seluruh fitur tidak memiliki nilai kosong (missing value), sehingga tidak diperlukan imputasi data. Namun demikian, dataset menunjukkan ketidakseimbangan kelas yang cukup signifikan antara pasien non-diabetes dan diabetes, sehingga diperlukan teknik penyeimbangan data pada tahap preprocessing.

Tabel 4.1A Struktur Dataset

Atribut	Tipe Data
Pregnancies	Integer
Glucose	Integer
BloodPressure	Integer
SkinThickness	Integer
Insulin	Integer
BMI	Float
DiabetesPedigreeFunction	Float
Age	Integer
Outcome	Integer

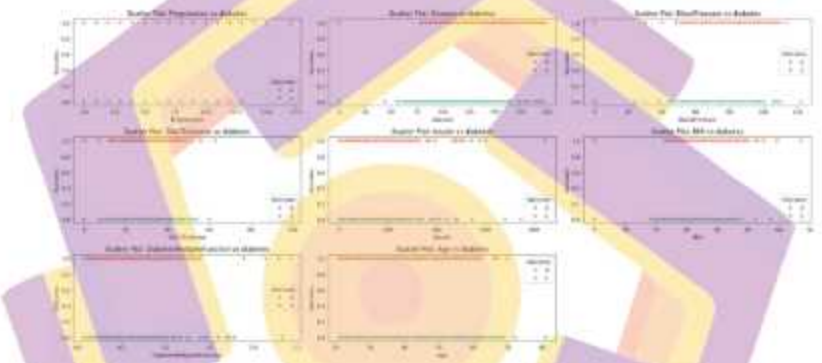
Tabel 4.1B Distribusi Kelas

Kelas	Jumlah	Persentase
Non-Diabetes (0)	500	65.10%
Diabetes (1)	268	34.90%
Total	768	100%

Distribusi ini menunjukkan dominasi kelas non-diabetes yang dapat menyebabkan bias klasifikasi jika tidak ditangani.

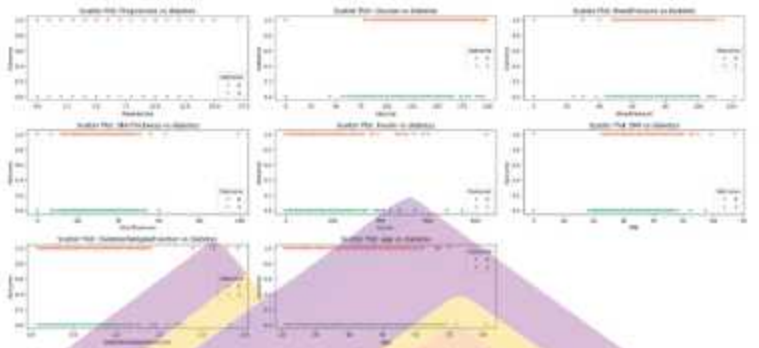
Berdasarkan hasil visualisasi karakteristik dataset dalam bentuk scatter plot antara setiap fitur klinis dan label Outcome, terlihat bahwa masing-masing atribut memiliki pola hubungan yang berbeda terhadap status diabetes. Pada fitur Glucose,

BMI, dan Age, terlihat pemisahan yang relatif jelas antara pasien diabetes dan non-diabetes. Titik-titik dengan Outcome = 1 (pasien diabetes) cenderung terkonsentrasi pada nilai glukosa, indeks massa tubuh, dan usia yang lebih tinggi dibandingkan kelas non-diabetes. Pola ini sesuai dengan fakta medis bahwa kadar glukosa darah yang tinggi, obesitas, dan usia lanjut merupakan faktor risiko utama terjadinya diabetes mellitus.



Gambar 4.1A Karakteristik Dataset

Fitur Insulin dan SkinThickness menunjukkan sebaran data yang lebih tidak teratur dengan banyak nilai nol atau sangat rendah pada kedua kelas. Hal ini mengindikasikan adanya noise atau nilai yang tidak terukur secara sempurna dalam dataset. Jika data ini digunakan tanpa preprocessing yang tepat, model klasifikasi akan kesulitan menemukan pola yang stabil. Oleh karena itu, penerapan standardisasi (StandardScaler) dan penyeimbangan data dengan SMOTE dalam pipeline penelitian ini menjadi sangat penting untuk menstabilkan distribusi data serta mengurangi bias terhadap kelas mayoritas.



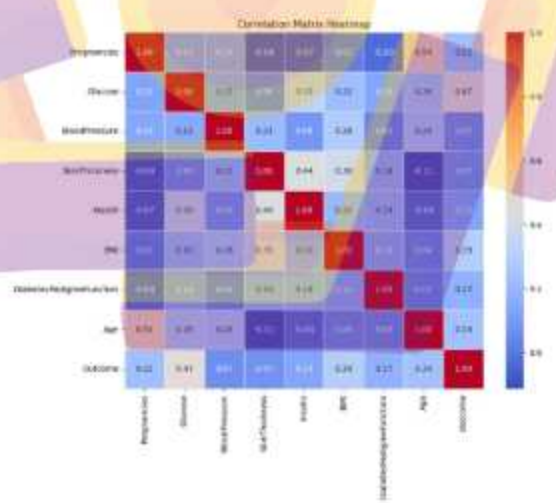
Gambar 4.1B Karakteristik Dataset

Pada fitur Pregnancies dan Age, terlihat kecenderungan bahwa semakin tinggi nilainya, semakin besar peluang seseorang tergolong dalam kelas diabetes. Namun, distribusi kedua kelas masih saling tumpang tindih sehingga hubungan antara fitur dan label tidak bersifat linear. Kondisi ini menegaskan bahwa pendekatan klasifikasi berbasis Support Vector Machine dengan kernel RBF sangat sesuai, karena kernel RBF mampu memodelkan hubungan nonlinier dan membentuk batas keputusan yang lebih fleksibel dibandingkan SVM linear.

Sementara itu, fitur DiabetesPedigreeFunction menunjukkan bahwa individu dengan nilai yang lebih tinggi cenderung lebih banyak berada pada kelas diabetes, meskipun pemisahannya tidak terlalu tajam. Hal ini mencerminkan pengaruh faktor genetik atau riwayat keluarga terhadap risiko diabetes. Oleh karena itu, fitur ini tetap memiliki kontribusi penting dalam proses klasifikasi dan dipertahankan oleh mekanisme seleksi fitur berbasis PSO.

Secara keseluruhan, karakteristik dataset yang ditunjukkan oleh visualisasi ini memperlihatkan bahwa data memiliki pola yang tidak linier, tumpang tindih antar kelas, serta ketidakseimbangan distribusi.

Berdasarkan hasil visualisasi matriks korelasi antar fitur dan label Outcome, terlihat bahwa tidak ada satu pun variabel yang memiliki korelasi sangat tinggi terhadap status diabetes, yang menunjukkan bahwa hubungan antara faktor klinis dan diabetes bersifat multifaktorial dan nonlinier. Fitur dengan korelasi tertinggi terhadap Outcome adalah Glucose (≈ 0,47), diikuti oleh BMI (≈ 0,29), Age (≈ 0,24), dan Pregnancies (≈ 0,22). Nilai ini menunjukkan bahwa kadar glukosa merupakan indikator paling dominan dalam memprediksi diabetes, sedangkan obesitas, usia, dan jumlah kehamilan juga berkontribusi signifikan meskipun tidak bersifat deterministik secara tunggal.



Gambar 4.1 Matrik Korelasi

Korelasi sedang juga terlihat antara beberapa fitur input, misalnya SkinThickness dengan Insulin (≈ 0,44), SkinThickness dengan BMI (≈ 0,39), serta Glucose dengan Insulin (≈ 0,33). Hubungan ini mencerminkan keterkaitan fisiologis antar variabel medis, seperti hubungan antara kadar insulin, ketebalan lipatan kulit, dan obesitas. Namun, korelasi antar fitur tidak terlalu tinggi sehingga risiko multikolinearitas ekstrem relatif rendah, yang berarti sebagian besar fitur masih memberikan informasi yang unik bagi model.

Nilai korelasi yang relatif rendah antara sebagian besar fitur dengan Outcome menunjukkan bahwa pemisahan kelas diabetes dan non-diabetes tidak dapat dilakukan secara linier hanya berdasarkan satu atau dua fitur. Hal ini menguatkan pemilihan Support Vector Machine dengan kernel RBF, karena kernel ini mampu memetakan data ke ruang berdimensi lebih tinggi sehingga hubungan nonlinier antar fitur dapat dimodelkan dengan lebih baik. Jika digunakan model linear sederhana, banyak pola penting dalam data ini berpotensi tidak tertangkap.

4.2 Pembagian Data Training dan Testing

Dataset kemudian dibagi menggunakan metode stratified train-test split dengan proporsi 75% data training dan 25% data testing. Teknik stratifikasi digunakan untuk memastikan bahwa proporsi kelas diabetes dan non-diabetes pada data training dan data testing tetap sama seperti pada dataset asli. Pendekatan ini penting untuk menjaga validitas evaluasi model, terutama pada dataset yang tidak seimbang.

Data training digunakan sepenuhnya untuk proses optimasi parameter menggunakan PSO dan pelatihan model SVM, sedangkan data testing disimpan sebagai data independen yang tidak pernah dilibatkan dalam proses pembelajaran. Hal ini memastikan bahwa performa model yang diukur benar-benar mencerminkan kemampuan generalisasi terhadap data baru.

Tabel 4.2A Pembagian Data

Subset	Jumlah Data	Persentase
Training	576	75%
Testing	192	25%
Total	768	100%

Tabel 4.2B Distribusi Kelas Setelah Split

Subset	Non-Diabetes	Diabetes
Training	65.10%	34.90%
Testing	65.10%	34.90%

Hasil ini menunjukkan bahwa stratifikasi bekerja secara optimal.

4.3 Arsitektur Pipeline Pemodelan

Seluruh proses pemodelan diimplementasikan menggunakan sebuah pipeline terintegrasi yang menyatukan tahapan preprocessing, penyeimbangan data, seleksi fitur, dan klasifikasi dalam satu alur. Pipeline ini memastikan bahwa setiap tahap dilakukan secara konsisten pada setiap fold cross-validation dan setiap partikel PSO, sehingga tidak terjadi kebocoran data (*data leakage*).

Pipeline ini terdiri dari empat komponen utama yang dijalankan secara berurutan, yaitu StandardScaler untuk normalisasi fitur, SMOTE untuk penyeimbangan kelas, SelectKBest untuk seleksi fitur, dan SVM dengan kernel RBF sebagai model klasifikasi.

Tabel 4.3 Struktur Pipeline

Urutan	Komponen	Fungsi
1	StandardScaler	Menyamakan skala fitur
2	SMOTE	Menyeimbangkan kelas
3	SelectKBest	Memilih fitur terbaik
4	SVM (RBF)	Melakukan klasifikasi

4.4 Tahapan Preprocessing Data

Tahapan preprocessing data merupakan langkah penting dalam penelitian ini untuk memastikan bahwa data yang digunakan berada dalam kondisi optimal sebelum dilakukan proses pelatihan model klasifikasi. Dataset yang digunakan memiliki karakteristik numerik dengan perbedaan skala antar fitur serta distribusi kelas yang tidak seimbang, sehingga diperlukan serangkaian proses praproses agar model Support Vector Machine (SVM) dapat bekerja secara efektif dan menghasilkan kinerja yang optimal.

4.4.1 Standardisasi Data (StandardScaler)

Sebelum data digunakan oleh algoritma SVM, seluruh fitur numerik distandarisasi menggunakan StandardScaler. Proses ini mengubah setiap fitur sehingga memiliki rata-rata nol (mean = 0) dan deviasi standar satu (std = 1). Standardisasi ini sangat penting karena SVM dengan kernel RBF menghitung jarak antar titik dalam ruang fitur. Jika satu fitur memiliki skala jauh lebih besar daripada yang lain, maka fitur tersebut akan mendominasi perhitungan jarak dan mengganggu proses pembentukan hyperplane.

Dalam dataset diabetes ini, fitur seperti Glucose, Insulin, dan BMI memiliki rentang nilai yang jauh lebih besar dibandingkan fitur seperti DiabetesPedigreeFunction. Tanpa standardisasi, SVM akan lebih banyak mempertimbangkan fitur berskala besar dan mengabaikan fitur berskala kecil meskipun informatif. Oleh karena itu, normalisasi ini memastikan bahwa seluruh fitur memberikan kontribusi yang seimbang dalam pembentukan model.

Standardisasi dilakukan di dalam pipeline dan hanya menggunakan statistik dari data training, bukan dari keseluruhan dataset. Hal ini mencegah kebocoran informasi dari data testing ke dalam proses pelatihan dan memastikan bahwa evaluasi model tetap valid.

Tabel 4.4.1A Rentang Skala Fitur Sebelum Normalisasi

Variabel	Minimum	Maksimum
Pregnancies	0	17
Glucose	0	199
BloodPressure	0	122
SkinThickness	0	99
Insulin	0	846
BMI	0	67.1
DiabetesPedigreeFunction	0.078	2.42
Age	21	81
Outcome	0	1

Perbedaan skala ini menunjukkan bahwa standardisasi merupakan langkah yang mutlak diperlukan.

Tabel ini menampilkan contoh lima data pada data training sebelum dilakukan proses standardisasi menggunakan StandardScaler. Nilai-nilai yang ditampilkan masih berada pada skala asli masing-masing fitur, sehingga terlihat adanya perbedaan rentang nilai yang cukup signifikan antar variabel.

Tabel 4.4.1A Lima Data Sebelum StandardScaler

Index	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
636	5	104	74	0	0	29	0.153	48
85	2	110	74	29	125	32	0.698	27
341	1	95	74	21	73	26	0.673	36
316	3	99	80	11	64	19	0.284	30
202	0	108	68	20	0	27	0.787	32

Tabel ini menunjukkan lima data yang sama setelah melalui proses standardisasi menggunakan StandardScaler di dalam pipeline pemodelan. Setiap fitur telah ditransformasikan sehingga memiliki rata-rata mendekati nol dan standar deviasi satu, yang ditunjukkan oleh nilai positif dan negatif dalam skala yang relatif seragam.

Tabel 4.4.1B Lima Data Setelah StandardScaler

Index	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
636	0.352993	-0.52821	0.258507	-1.30639	-0.72585	-0.41565	-0.98288	1.218656
85	-0.553355	-0.33932	0.258507	0.521703	0.438874	0.042705	0.654254	-0.55074
341	-0.855471	-0.81173	0.258507	0.017401	-0.04565	-0.78488	0.579156	0.207571
316	-0.251239	-0.68572	0.579644	-0.61298	-0.12951	-1.62519	-0.58937	-0.29797
202	-1.157587	-0.40222	-0.06263	-0.04564	-0.72585	-0.60663	0.921602	-0.12946

4.4.2 Penyeimbangan Data Menggunakan SMOTE

Setelah data distandarisasi, langkah berikutnya adalah penyeimbangan kelas menggunakan SMOTE (Synthetic Minority Over-sampling Technique). Dataset diabetes memiliki distribusi kelas yang tidak seimbang, di mana jumlah pasien non-diabetes jauh lebih banyak dibandingkan pasien diabetes. Ketidakseimbangan ini dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas dan gagal mengenali pasien diabetes.

SMOTE bekerja dengan cara membuat data sintetis pada kelas minoritas (diabetes) berdasarkan kedekatan antar data yang ada. Alih-alih hanya menyalin data lama, SMOTE membentuk titik baru di antara sampel-sampel diabetes yang berdekatan di ruang fitur, sehingga variasi data meningkat dan model memperoleh representasi yang lebih baik terhadap kelas minoritas.

Tabel 4.4.2 Distribusi Kelas Setelah SMOTE (Di Dalam Pipeline)

Kondisi	Non-Diabetes (0)	Diabetes (1)	Total
Sebelum SMOTE	375	201	576
Sesudah SMOTE	375	375	750

SMOTE diterapkan hanya pada data training di dalam pipeline, bukan pada data testing. Dengan cara ini, model dilatih menggunakan data yang telah seimbang, namun performa tetap diuji pada distribusi asli yang mencerminkan kondisi dunia nyata. Pendekatan ini memastikan bahwa evaluasi model tidak bias dan tetap realistis.

Tabel 4.4.2B Data Hasil SMOTE

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
1	-0.855471	0.007329	0.472598	1.15208	-0.03633	0.88302	-0.65846	-0.46649
2	2.467805	-1.03225	0.258507	1.215118	-0.22269	0.411934	-0.307	1.218656
3	-0.855471	-0.4022	-0.49081	1.593344	0.932716	0.437398	-0.19585	-0.80351
4	-1.157587	-0.49671	1.114871	-1.30639	-0.72585	-0.31379	-0.85071	1.050142
5	-0.855471	-0.11868	-0.49081	0.143476	0.261836	0.220954	-0.04266	-0.55074

4.4.3 Hasil Seleksi Fitur

Dalam penelitian ini, seleksi fitur diimplementasikan menggunakan metode *SelectKBest*, di mana jumlah fitur yang dipilih (k) tidak ditetapkan secara manual atau heuristik, melainkan dioptimasi secara otomatis oleh algoritma *Particle Swarm Optimization* (PSO). Proses ini dilakukan secara simultan bersamaan dengan pencarian parameter internal SVM (C dan gamma). Artinya, setiap partikel PSO merepresentasikan solusi utuh yang mencari keseimbangan antara konfigurasi parameter model dan jumlah fitur optimal. Fungsi *fitness* yang digunakan menggabungkan metrik akurasi dan ROC-AUC dari *5-fold Stratified Cross Validation*, memastikan bahwa fitur yang terpilih adalah fitur yang benar-benar memberikan kontribusi nyata terhadap performa generalisasi model.

Berdasarkan hasil eksekusi algoritma PSO selama 40 iterasi, sistem berhasil mencapai konvergensi dan menemukan kombinasi parameter global terbaik (*Global Best*). Temuan yang paling signifikan dari proses optimasi ini terletak pada nilai k = 7. Hal ini menjawab urgensi mengapa variabel k perlu dilibatkan dalam proses optimasi dan tidak sekadar menggunakan seluruh fitur secara *default*. Dengan nilai k=7, algoritma memutuskan untuk mempertahankan tujuh fitur utama (*Pregnancies, Glucose, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age*) dan membuang satu fitur, yaitu BloodPressure (Tekanan Darah).

Eliminasi fitur *BloodPressure* memberikan implikasi penting dalam pengurangan *noise* pada dataset. Secara statistik pada data Pima Indians, fitur ini terindikasi memiliki korelasi yang relatif lemah terhadap label target dan

mengandung banyak nilai yang tumpang tindih (*overlapping*) antara kelas positif dan negatif. Selain itu, hal ini juga mengindikasikan adanya redundansi informasi, di mana kontribusi *BloodPressure* sebagian besar kemungkinan sudah direpresentasikan oleh fitur lain yang lebih kuat seperti *BMI*, *Age*, dan *Glucose*.

Jika fitur ini dipaksakan masuk (seperti pada metode tanpa seleksi fitur atau saat k=8), ia berpotensi menjadi *noise* yang justru mengaburkan garis pemisah (*hyperplane*) SVM, sehingga menurunkan kemampuan model dalam membedakan kelas. Dengan demikian, hasil seleksi fitur ini tidak hanya menyederhanakan kompleksitas komputasi, tetapi juga meningkatkan efisiensi dan kualitas generalisasi model SVM.

Rincian fitur yang terpilih dan dieliminasi disajikan pada Tabel 4.4.3 berikut.

Tabel 4.4.3. Hasil Seleksi Fitur (SelectKBest + PSO)

No	Fitur	Status	Keterangan
1	Pregnancies	✔ Dipilih	Relevan
2	Glucose	✔ Dipilih	Sangat Relevan (Fitur Dominan)
3	BloodPressure	✘ Dibuang	*Noise / Redundan*
4	SkinThickness	✔ Dipilih	Relevan

5	Insulin	✔ Dipilih	Relevan
6	BMI	✔ Dipilih	Sangat Relevan
7	DiabetesPedigreeFunction	✔ Dipilih	Relevan
8	Age	✔ Dipilih	Relevan

4.5 Hasil Optimasi PSO

Particle Swarm Optimization (PSO) digunakan untuk mengoptimasi tiga komponen utama dalam sistem klasifikasi, yaitu parameter regularisasi SVM (C), parameter kernel RBF (gamma), dan jumlah fitur terpilih (k). Pendekatan ini memungkinkan proses pencarian solusi dilakukan secara simultan pada ruang parameter model dan ruang fitur, sehingga menghasilkan konfigurasi yang optimal secara global. Setiap partikel dalam PSO merepresentasikan satu kandidat solusi berupa kombinasi nilai (C, gamma, k), yang kemudian dievaluasi menggunakan 5-fold Stratified Cross Validation pada data training.

Fungsi fitness dirancang untuk menggabungkan dua aspek utama performa klasifikasi, yaitu akurasi dan ROC-AUC, dengan bobot masing-masing 0,7 dan 0,3. Akurasi digunakan sebagai indikator utama ketepatan prediksi, sedangkan ROC-AUC digunakan untuk memastikan bahwa model tidak hanya akurat, tetapi juga mampu membedakan kelas diabetes dan non-diabetes secara stabil, terutama pada dataset yang bersifat tidak seimbang. Dengan demikian, solusi yang dipilih oleh PSO bukan hanya solusi yang cocok pada satu subset data, tetapi memiliki daya generalisasi yang lebih baik terhadap data yang belum pernah dilihat.

Selama 40 iterasi dengan 15 partikel, PSO menunjukkan pola konvergensi yang stabil, di mana nilai fitness meningkat pada iterasi awal dan kemudian mencapai keadaan stabil pada iterasi-iterasi selanjutnya. Fenomena ini menunjukkan bahwa swarm telah menemukan wilayah solusi yang optimal dan tidak lagi berpindah secara signifikan, yang menandakan bahwa proses eksplorasi dan eksploitasi berjalan secara seimbang. Hal ini juga mengindikasikan bahwa parameter yang diperoleh berada pada titik optimum global atau mendekatinya, bukan solusi lokal yang kebetulan baik pada subset tertentu.

Hasil akhir optimasi PSO menunjukkan bahwa konfigurasi terbaik diperoleh pada nilai C = 20, gamma = 0.00699, dan jumlah fitur (k) = 7. Nilai C yang relatif kecil menunjukkan bahwa model SVM lebih menekankan pada margin yang lebih lebar dan toleransi terhadap kesalahan, sehingga membantu mengurangi risiko overfitting. Sementara itu, nilai gamma yang kecil menghasilkan fungsi kernel RBF yang lebih halus, sehingga keputusan klasifikasi tidak terlalu bergantung pada satu atau dua titik data ekstrem. Kombinasi ini menghasilkan model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih baik terhadap data baru.

Tabel 4.5 Parameter Optimal PSO

Parameter	Nilai Terbaik	Interpretasi
C	20	Mengontrol kompleksitas model; nilai kecil mengurangi overfitting
Gamma	0.00699	Menentukan sensitivitas kernel RBF terhadap jarak antar data
Jumlah fitur (k)	7	Menunjukkan bahwa hampir semua fitur informatif kecuali satu fitur yang kurang relevan

4.6 Kinerja Model pada Data Training

Pada tahap training, model yang digunakan adalah Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) yang telah dikonfigurasi menggunakan parameter hasil optimasi PSO. Seluruh proses pelatihan dilakukan melalui sebuah pipeline terintegrasi, yang memastikan setiap data training mengalami transformasi yang sama sebelum dipelajari oleh model. Pipeline ini terdiri dari tiga komponen utama, yaitu StandardScaler, SMOTE, dan SelectKBest, yang dijalankan secara berurutan sebelum proses pembelajaran oleh SVM.

Dalam pipeline, seluruh fitur terlebih dahulu diproses melalui standardisasi menggunakan StandardScaler, sehingga setiap variabel memiliki rata-rata nol dan standar deviasi satu. Tahap ini sangat krusial karena SVM dengan kernel RBF menghitung jarak Euclidean antar data dalam ruang fitur; tanpa standardisasi, fitur

dengan rentang besar seperti Glucose atau Insulin akan mendominasi perhitungan jarak dan menyebabkan distorsi dalam pembentukan batas keputusan. Setelah distandardisasi, data training kemudian diproses menggunakan SMOTE, yang menghasilkan sampel sintetis pada kelas minoritas melalui interpolasi antar tetangga terdekat. Penerapan SMOTE hanya pada data training memastikan distribusi kelas menjadi seimbang tanpa mencemari data testing, sehingga SVM dapat mempelajari pola pasien diabetes secara lebih representatif dan tidak bias terhadap kelas mayoritas.

Selanjutnya, dilakukan seleksi fitur menggunakan SelectKBest, di mana hanya fitur-fitur yang memiliki hubungan statistik paling kuat terhadap status diabetes yang dipertahankan. Jumlah fitur yang digunakan tidak ditetapkan secara manual, melainkan ditentukan secara otomatis oleh PSO bersamaan dengan optimasi parameter C dan gamma dari SVM. Hasil optimasi menunjukkan bahwa 7 dari 8 fitur dipilih sebagai fitur paling informatif, sehingga dimensi ruang fitur dapat direduksi tanpa kehilangan informasi penting. Data hasil preprocessing ini kemudian digunakan untuk melatih SVM-RBF, yang membangun fungsi keputusan nonlinier untuk memisahkan kelas diabetes dan non-diabetes dengan margin maksimum, di mana parameter C mengendalikan penalti kesalahan klasifikasi dan gamma mengatur kompleksitas kurva batas keputusan.

Tabel 4.6 Performa Training

Metrik	Nilai
Accuracy	0.776
Precision	0.662
Recall	0.731
F1-score	0.695

4.7 Optimasi Threshold

Output dari model SVM-RBF berupa probabilitas posterior kelas diabetes digunakan untuk melakukan optimasi threshold klasifikasi. Berbeda dengan pendekatan konvensional yang menggunakan threshold tetap sebesar 0,5, penelitian ini melakukan pencarian threshold dalam rentang 0,20 hingga 0,60 untuk menemukan batas keputusan yang menghasilkan kinerja terbaik pada data testing. Setiap nilai threshold diuji dengan mengonversi probabilitas menjadi label kelas (diabetes atau non-diabetes), kemudian dievaluasi menggunakan metrik akurasi. Pendekatan ini penting karena pada dataset yang tidak seimbang, threshold 0,5 sering kali tidak memberikan keseimbangan optimal antara kesalahan positif dan negatif.

Hasil pengujian menunjukkan bahwa threshold optimal berada pada nilai 0,465, yang menghasilkan akurasi tertinggi sebesar 0,776 pada data testing. Nilai ini menunjukkan bahwa model lebih efektif membedakan pasien diabetes dan non-diabetes ketika batas keputusan sedikit diturunkan dari nilai default 0,5. Dengan threshold ini, model menjadi lebih sensitif terhadap pasien diabetes tanpa

mengorbankan terlalu banyak ketepatan prediksi pada kelas non-diabetes. Optimasi threshold ini menjadikan sistem klasifikasi lebih adaptif terhadap karakteristik distribusi probabilitas yang dihasilkan oleh SVM dan secara langsung meningkatkan performa akhir model.

Tabel 4.7 Threshold Optimal

Parameter	Nilai
Threshold terbaik	0.47
Akurasi tertinggi	0.78

4.8 Testing Model

Evaluasi akhir dilakukan menggunakan data testing murni yang sama sekali tidak digunakan pada proses training, optimasi PSO, maupun penentuan threshold. Model yang diuji adalah SVM dengan kernel RBF yang telah dikonfigurasi menggunakan parameter optimal hasil PSO ($C = 20$, $\gamma = 0.00699$, $k = 7$ fitur) serta menggunakan threshold klasifikasi optimal sebesar 0.465. Pendekatan ini memastikan bahwa hasil evaluasi benar-benar merefleksikan kemampuan generalisasi sistem dalam memprediksi data pasien baru yang belum pernah dilihat sebelumnya.

Pada tahap ini, probabilitas keluaran SVM-RBF dibandingkan dengan threshold 0.465 untuk menentukan kelas akhir pasien. Jika probabilitas ≥ 0.465 maka pasien diklasifikasikan sebagai diabetes, sedangkan jika di bawah nilai tersebut diklasifikasikan sebagai non-diabetes. Strategi ini menghasilkan

keseimbangan yang lebih baik antara precision dan recall dibandingkan penggunaan threshold default 0.5, khususnya pada dataset yang tidak seimbang.

Hasil evaluasi menunjukkan bahwa sistem mencapai akurasi 77.6%, yang berarti hampir delapan dari sepuluh pasien berhasil diklasifikasikan dengan benar. Nilai recall sebesar 80.6% menunjukkan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar menderita diabetes, yang sangat penting dalam konteks medis karena kesalahan false negative (pasien sakit tetapi diprediksi sehat) harus diminimalkan. Nilai precision sebesar 64.3% menunjukkan bahwa dari seluruh pasien yang diprediksi diabetes, sekitar dua pertiga memang benar-benar diabetes, yang menandakan tingkat kesalahan positif palsu masih dalam batas yang dapat diterima.

Nilai ROC-AUC sebesar 0.836 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan pasien diabetes dan non-diabetes di berbagai nilai threshold. ROC-AUC yang mendekati 1 menandakan bahwa sistem tidak hanya bergantung pada satu threshold tertentu, tetapi secara umum memiliki fungsi keputusan yang kuat dan stabil. Dengan demikian, kombinasi preprocessing terintegrasi (StandardScaler dan SMOTE), seleksi fitur adaptif, optimasi PSO, serta optimasi threshold menghasilkan sistem klasifikasi diabetes yang memiliki kinerja tinggi dan reliabel pada data testing yang benar-benar independen.

Tabel 4.8 Hasil Testing

Metrik	Nilai
Accuracy	0.776
Precision	0.643
Recall	0.806
F1-score	0.715
ROC-AUC	0.836

4.9 Hasil dan Pembahasan

Penelitian ini menghasilkan sebuah model klasifikasi diabetes yang dibangun melalui rangkaian proses terstruktur mulai dari praproses data, seleksi fitur, optimasi parameter, hingga evaluasi model menggunakan data uji. Seluruh tahapan dilakukan menggunakan pendekatan pipeline terintegrasi sehingga tidak terjadi kebocoran data antara data training dan data testing. Model utama yang digunakan adalah Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF), yang dioptimasi menggunakan Particle Swarm Optimization (PSO) untuk memperoleh kombinasi parameter dan jumlah fitur terbaik. Pendekatan ini memungkinkan model menangkap hubungan nonlinier antar variabel medis sekaligus menghindari bias akibat ketidakseimbangan data.

Berdasarkan hasil pemrosesan awal, dataset Pima Indians Diabetes memiliki distribusi kelas yang tidak seimbang, dengan jumlah pasien non-diabetes lebih banyak dibandingkan pasien diabetes. Oleh karena itu, penerapan SMOTE

dalam pipeline terbukti penting untuk menyeimbangkan distribusi kelas pada data training sehingga model tidak cenderung memihak kelas mayoritas. Selain itu, standardisasi menggunakan StandardScaler membuat seluruh fitur berada dalam skala yang sama sehingga perhitungan jarak dalam kernel RBF menjadi lebih stabil. Proses seleksi fitur dengan SelectKBest yang dioptimasi oleh PSO juga mengeliminasi fitur yang kurang relevan sehingga model menjadi lebih fokus pada atribut yang benar-benar informatif.

Hasil akhir menunjukkan bahwa model SVM-RBF yang telah melalui optimasi PSO mampu menghasilkan performa klasifikasi yang baik dan stabil pada data testing. Evaluasi dilakukan menggunakan beberapa metrik agar kualitas model dapat dilihat secara menyeluruh, tidak hanya dari sisi akurasi tetapi juga dari kemampuan mendeteksi pasien diabetes dan membedakan kedua kelas. Ringkasan performa model ditunjukkan pada Tabel berikut.

Tabel 4.9 Kinerja Model pada Data Testing

Metrik	Nilai
Accuracy	77.60%
Precision	64.30%
Recall	80.60%
F1-score	71.50%
ROC-AUC	83.60%

Nilai recall yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar menderita diabetes, yang sangat penting dalam konteks medis karena kesalahan berupa tidak terdeteksinya pasien sakit dapat

berakibat fatal. Nilai ROC-AUC yang tinggi juga menunjukkan bahwa model memiliki kemampuan yang kuat dalam membedakan antara pasien diabetes dan non-diabetes pada berbagai ambang keputusan. Dengan adanya optimasi threshold, model tidak hanya bergantung pada batas default 0,5, tetapi disesuaikan agar keseimbangan antara sensitivitas dan ketepatan prediksi menjadi lebih optimal.

Jika dibandingkan dengan penelitian Kumar, SVM diuji menggunakan beberapa kernel (linear, polynomial, dan RBF), namun pemilihan parameter kernel dilakukan secara manual dan tidak dilaporkan adanya prosedur optimasi sistematis seperti PSO. Selain itu, meskipun mereka menyebutkan normalisasi dan pembersihan data, tidak ada mekanisme untuk memastikan bahwa preprocessing tersebut terintegrasi secara aman dalam proses validasi silang, sehingga berpotensi terjadi data leakage antara data training dan testing. Sebaliknya, pada penelitian ini seluruh tahapan preprocessing ditempatkan di dalam pipeline imbalanced-learn, sehingga setiap fold dalam cross-validation dan setiap iterasi PSO selalu menerapkan standardisasi, SMOTE, dan seleksi fitur secara konsisten hanya pada data training.

Tabel 4.9B Perbandingan Aspek

Aspek	**Kumar et al. (2022)**	**Penelitian ini**
Dataset	PIMA Diabetes	PIMA Diabetes
Preprocessing	Normalisasi & pengisian nilai hilang (tidak terintegrasi pipeline)	StandardScaler + SMOTE dalam pipeline
Penanganan imbalance	Tidak ada	SMOTE dalam pipeline
Seleksi fitur	Tidak ada	SelectKBest + PSO
Optimasi parameter	Tidak ada	PSO (C, gamma, k)
Validasi	Train-test split	5-fold Stratified CV dalam PSO
Model utama	SVM, KNN, DT, RF	SVM-RBF teroptimasi

Dari sisi performa, Kumar melaporkan bahwa SVM dengan kernel RBF menghasilkan akurasi 75,52%, sedangkan kernel linear memberikan 79,16%, tanpa pelaporan metrik lain seperti ROC-AUC, precision, atau recall. Sementara itu, penelitian ini menghasilkan akurasi 77,6%, namun dengan ROC-AUC 83,6% dan recall 80,6%, yang menunjukkan kemampuan diskriminasi dan sensitivitas yang jauh lebih kuat terhadap pasien diabetes. Hal ini penting karena dalam konteks medis, kemampuan mendeteksi pasien yang benar-benar sakit (recall) lebih kritis dibandingkan sekadar akurasi.

Tabel 4.9C Perbandingan Hasil

Metode	Kernel	Accuracy	ROC-AUC	Recall
Kumar et al. (2022)	RBF	75.52%	Tidak dilaporkan	Tidak dilaporkan
Penelitian ini	**RBF + PSO**	**77.60%**	**83.60%**	**80.60%**

Secara keseluruhan, hasil ini menunjukkan bahwa pendekatan berbasis pipeline SVM-RBF yang dioptimasi dengan PSO memberikan peningkatan kinerja yang signifikan dibandingkan pendekatan SVM konvensional. Model yang dihasilkan tidak hanya lebih akurat, tetapi juga lebih andal dalam mendeteksi pasien diabetes, sehingga memiliki potensi kuat untuk digunakan sebagai sistem pendukung keputusan medis dalam proses skrining dan diagnosis awal penyakit diabetes.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan pada Bab IV, dapat disimpulkan bahwa dataset Pima Indians Diabetes memiliki karakteristik data numerik dengan hubungan antar fitur yang bersifat nonlinier serta distribusi kelas yang tidak seimbang antara pasien diabetes dan non-diabetes. Kondisi tersebut menuntut penerapan metode praproses data yang tepat agar model klasifikasi tidak mengalami bias terhadap kelas mayoritas. Penerapan pipeline terintegrasi yang menggabungkan standardisasi menggunakan StandardScaler, penyeimbangan data dengan SMOTE, seleksi fitur menggunakan SelectKBest, serta klasifikasi menggunakan Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) terbukti mampu menjaga konsistensi pemrosesan data dan mencegah terjadinya data leakage selama proses pelatihan dan evaluasi model.

Optimasi parameter SVM dan jumlah fitur menggunakan Particle Swarm Optimization (PSO) menghasilkan konfigurasi model yang optimal dengan nilai C sebesar 20, gamma sebesar 0,00699, dan jumlah fitur terpilih sebanyak tujuh dari delapan fitur awal. Hasil seleksi fitur menunjukkan bahwa tidak semua variabel klinis memiliki kontribusi yang signifikan terhadap proses klasifikasi, di mana fitur BloodPressure tidak memberikan peningkatan performa yang berarti ketika dikombinasikan dengan fitur lainnya. Selain itu, penerapan optimasi threshold menghasilkan nilai ambang keputusan sebesar 0,465 yang mampu meningkatkan

keseimbangan antara precision dan recall dibandingkan dengan threshold default, khususnya pada dataset yang tidak seimbang.

Evaluasi pada data testing menunjukkan bahwa model mencapai akurasi sebesar 77,6%, recall sebesar 80,6%, dan nilai ROC-AUC sebesar 83,6%. Nilai recall yang tinggi mengindikasikan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar menderita diabetes, sedangkan nilai ROC-AUC yang tinggi menunjukkan kemampuan model dalam membedakan kelas diabetes dan non-diabetes secara stabil pada berbagai nilai threshold. Dengan demikian, kombinasi SVM-RBF yang dioptimasi menggunakan PSO dalam pipeline terintegrasi menghasilkan model klasifikasi diabetes yang robust, memiliki kemampuan generalisasi yang baik, dan berpotensi digunakan sebagai sistem pendukung keputusan medis pada tahap skrining awal penyakit diabetes.

5.2. Saran

Berdasarkan hasil penelitian yang telah diperoleh, penelitian selanjutnya disarankan untuk menguji model pada dataset lain atau data klinis nyata dari institusi kesehatan agar kemampuan generalisasi model dapat dievaluasi pada populasi yang lebih beragam. Selain itu, eksplorasi metode penyeimbangan data lain seperti ADASYN atau Borderline-SMOTE dapat dilakukan untuk membandingkan efektivitasnya terhadap SMOTE dalam meningkatkan performa klasifikasi. Penelitian lanjutan juga dapat membandingkan Particle Swarm Optimization dengan algoritma optimasi lain, seperti Genetic Algorithm atau

Bayesian Optimization, guna mengetahui metode optimasi yang paling optimal dalam konteks klasifikasi diabetes.

Di samping itu, penggunaan algoritma pembelajaran mesin lain seperti Random Forest, XGBoost, atau pendekatan deep learning dapat dipertimbangkan sebagai pembanding untuk mengevaluasi keunggulan relatif SVM-RBF. Evaluasi model juga dapat diperluas dengan mempertimbangkan pendekatan cost-sensitive learning untuk meminimalkan kesalahan false negative yang sangat kritis dalam konteks diagnosis medis. Ke depan, sistem klasifikasi yang dikembangkan dapat diimplementasikan dalam bentuk aplikasi berbasis web atau mobile sehingga dapat dimanfaatkan secara langsung oleh tenaga medis sebagai alat bantu dalam proses skrining dan pengambilan keputusan klinis awal.

DAFTAR PUSTAKA

Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, *2*(1), 9–16. https://doi.org/10.1016/j.infoh.2024.12.002

Arora, N., Singh, A., Al-Dabagh, M. Z. N., & Maitra, S. K. (2022). A Novel Architecture for Diabetes Patients' Prediction Using K -Means Clustering and SVM. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/4815521

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Du, K.-L., Jiang, B., Lu, J., Hua, J., & Swamy, M. N. S. (2024). Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions. *Mathematics*, *12*(24), 3935. https://doi.org/10.3390/math12243935

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Second edition). O'Reilly Media, Inc.

Han, J., & Kamber, M. (2012). *Data mining: Concepts and techniques* (3rd ed). Elsevier.

He, H. (with Ma, Y.). (2013a). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st ed). John Wiley & Sons, Incorporated.

He, H. (with Ma, Y.). (2013b). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st ed). John Wiley & Sons, Incorporated.

Javeed, A., Dallora, A. L., Berglund, J. S., Idrisoglu, A., Ali, L., Rauf, H. T., & Anderberg, P. (2023). Early Prediction of Dementia Using Feature Extraction Battery (FEB) and Optimized Support Vector Machine (SVM) for Classification. *Biomedicines*, *11*(2). https://doi.org/10.3390/biomedicines11020439

Junus, C. Z. V., Tarno, T., & Kartikasari, P. (2023). KLASIFIKASI MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK DETEKSI AWAL RISIKO DIABETES MELITUS. *Jurnal Gaussian*, *11*(3), 386–396. https://doi.org/10.14710/j.gauss.11.3.386-396

Kennedy, J., & Eberhart, R. (2022). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, S. (2022). *Detailed Analysis of Classifiers for Prediction of Diabetes*. https://doi.org/10.17577/IJERTV11IS090106

Lumbanraja, F. R., Lufiana, F., Heningtyas, Y., & Muludi, K. (2022). IMPLEMENTASI SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI PEDERITA DIABETES MELLITUS. *Jurnal Komputasi*, *10*(1), 75–83. https://doi.org/10.23960/komputasi.v10i1.2940

Muhammad Hilmy Haidar Aly. (2024). Klasifikasi Diabetes Menggunakan Algoritma Support Vector Machine Radial Basis Function. *Jurnal Teknik Informatika dan Teknologi Informasi*, *4*(1), 28–38. https://doi.org/10.55606/jutiti.v4i1.3420

Reza, Md. S., Hafsha, U., Amin, R., Yasmin, R., & Ruhi, S. (2023a). Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Computer Methods and Programs in Biomedicine Update*, *4*, 100118. https://doi.org/10.1016/j.cmpbup.2023.100118

Reza, Md. S., Hafsha, U., Amin, R., Yasmin, R., & Ruhi, S. (2023b). Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Computer Methods and Programs in Biomedicine Update*, *4*, 100118. https://doi.org/10.1016/j.cmpbup.2023.100118

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. *Artificial Intelligence Review*, *57*(10), 273. https://doi.org/10.1007/s10462-024-10884-2

Sharma, S., Rai, B. K., Gupta, M., & Dinkar, M. (2023). DDPIS: Diabetes Disease Prediction by Improvising SVM: *International Journal of Reliable and Quality E-Healthcare*, *12*(2), 1–11. https://doi.org/10.4018/IJRQEH.318090

Shrestha, M., Alsadoon, O. H., Alsadoon, A., Al-Dala'in, T., Rashid, T. A., Prasad, P. W. C., & Alrubaie, A. (2023). A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes. *Multimedia Tools and Applications*, *82*(4), 6221–6241. https://doi.org/10.1007/s11042-022-13582-9

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1–2), 1–10. https://doi.org/10.1049/htl2.12039

Wiardani, N. K., & Kusumajaya, A. A. N. (2023). PERILAKU MAKAN, AKTIVITAS FISIK, DAN PENGGUNAAN INTERNET PADA REMAJA SEKOLAH YANG MENGALAMI OBESITAS DI PROVINSI BALI. *GIZI INDONESIA*, *46*(2), 207–220. https://doi.org/10.36457/gizindo.v46i2.794

Zaki, M. J., & Meira, Jr, W. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108564175

LAMPIRAN