JURNAL DEVICE, VOL. 15 NO 2, November 2025, 220-232
ISSN: 0216-9185 | eISSN: 2746-8984

PENINGKATAN KINERJA KLASIFIKASI DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DENGAN KERNEL RADIAL BASIS FUNCTION (RBF)

Hasim As'ari [1)], Kusrini [2)]

[1) 2)] Universitas AMIKOM Yogyakarta
Email : hasimasari@students.amikom.ac.id [1)], kusrini@amikom.ac.id [2)]

ABSTRAK

Diabetes Mellitus merupakan salah satu penyakit kronis dengan prevalensi yang terus meningkat secara global. Deteksi dini terhadap penyakit ini sangat penting guna mengurangi risiko komplikasi yang lebih parah. Penelitian ini bertujuan untuk mengevaluasi performa algoritma *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF) dalam mengklasifikasikan penderita diabetes berdasarkan data medis. Untuk meningkatkan performa model, diterapkan berbagai tahapan *preprocessing* seperti normalisasi dengan *StandardScaler*, pembangkitan fitur non-linear dengan *PolynomialFeatures*, seleksi fitur dengan *SelectKBest*, serta penyeimbangan kelas menggunakan *ADASYN*. Dataset yang digunakan adalah Pima Indians Diabetes dari Kaggle, yang memiliki permasalahan ketidakseimbangan kelas. Hasil evaluasi menunjukkan bahwa model mampu mencapai nilai akurasi sebesar 76,6% dan nilai ROC AUC sebesar 0,861. Temuan ini menunjukkan bahwa pendekatan berbasis machine learning dengan pipeline yang tepat dapat menjadi solusi yang andal untuk mendukung deteksi dini Diabetes Mellitus secara otomatis.

Kata Kunci : *Diabetes Mellitus, SVM-RBF, ADASYN, Machine Learning, Klasifikasi, Preprocessing Data.*

ABSTRACT

Diabetes Mellitus is a chronic disease with a growing global prevalence. Early detection is essential to reduce the risk of severe complications. This study aims to evaluate the performance of the Support Vector Machine (SVM) algorithm with the Radial Basis Function (RBF) kernel in classifying diabetic patients based on medical data. To enhance model performance, several preprocessing steps were applied, including normalization using StandardScaler, non-linear feature generation using PolynomialFeatures, feature selection with SelectKBest, and class balancing using ADASYN. The dataset used was the Pima Indians Diabetes dataset from Kaggle, which presents class imbalance issues. The evaluation results show that the model achieved an accuracy of 76.6% and a ROC AUC score of 0.861. These findings indicate that a machine learning approach with a proper preprocessing pipeline can serve as a reliable tool for early automated detection of Diabetes Mellitus.

Keywords: Diabetes Mellitus, SVM-RBF, ADASYN, Machine Learning, Classification, Data Preprocessing.

1. PENDAHULUAN

Diabetes Mellitus merupakan penyakit metabolik kronis yang ditandai oleh peningkatan kadar glukosa darah akibat gangguan produksi atau fungsi insulin. Prevalensi diabetes melitus (DM) terus meningkat di seluruh dunia (Ruze et al., 2023). Menurut data dari *World Health Organization* (WHO), pada tahun 2021 lebih dari 422 juta orang di dunia hidup dengan diabetes, dan sebanyak 1,5 juta kematian terjadi setiap tahun akibat komplikasi yang berkaitan dengan penyakit ini (*World Health Organization, Diabetes*, 2021). Dampak dari diabetes tidak hanya dirasakan secara individual dalam bentuk penurunan kualitas hidup, tetapi juga berdampak secara sosial dan ekonomi, terutama di negara-negara berkembang. Diabetes Melitus juga menjadi penyebab utama kebutaan, penyakit jantung dan gagal ginjal (Kemenkes RI, 2021).

Di Indonesia, prevalensi diabetes juga menunjukkan peningkatan yang mengkhawatirkan. Data dari *International Diabetes Federation* (IDF) tahun 2019 mencatat bahwa Indonesia menempati peringkat ketujuh dunia dengan jumlah penderita mencapai 10,7 juta jiwa, atau sekitar 6,2% dari populasi dewasa. Kurangnya kesadaran masyarakat terhadap gaya hidup sehat serta rendahnya akses terhadap layanan deteksi dini menjadi penyebab utama tingginya angka penderita diabetes yang tidak terdiagnosis (IDF, l;o2019). Pemanfaatan teknologi seperti *machine learning* dalam bidang kesehatan menjadi salah satu pendekatan potensial yang dapat meningkatkan efektivitas diagnosis dan pemantauan penyakit ini secara lebih dini dan efisien (Thieme et al., 2020).

Penelitian oleh (Wong et al., n.d.) berfokus pada prediksi obesitas di kalangan pekerja dewasa Malaysia dengan memanfaatkan algoritma machine learning (ML) dan membandingkannya dengan metode Logistic Regression. Data penelitian diambil dari survei Malaysia's Healthiest Workplace tahun 2019 yang melibatkan 16.860 responden. Faktor yang dianalisis mencakup aspek demografis, karakteristik pekerjaan, persepsi terkait kesehatan dan berat badan, serta kebiasaan gaya hidup. Algoritma ML yang digunakan antara lain XGBoost, Random Forest (RF), dan Support Vector Machine (SVM), kemudian dibandingkan dengan Logistic Regression untuk memprediksi status kelebihan berat badan atau obesitas berdasarkan Indeks Massa Tubuh (IMT). Hasil evaluasi menunjukkan bahwa performa model ML relatif sebanding dengan Logistic Regression, dengan nilai AUC berkisar antara 0,78–0,81. XGBoost memberikan performa tertinggi dengan AUC 0,81, diikuti RF dan SVM dengan AUC 0,80, sementara Logistic Regression sedikit lebih rendah pada 0,78. SVM sendiri mencatat akurasi 72%, sensitivitas 65%, dan spesifisitas 77%, yang menandakan kinerja cukup baik dalam membedakan kelompok obesitas dan non-obesitas. Namun demikian, keterbatasan SVM terletak pada deteksi individu obesitas yang masih kurang optimal, sehingga disarankan adanya peningkatan performa melalui hyperparameter tuning serta penerapan teknik pra-pemrosesan tambahan. Hal ini relevan mengingat penelitian Wong hanya melakukan seleksi fitur, penanganan missing value, serta normalisasi dengan min-max scaling, namun belum menerapkan metode penyeimbangan data seperti SMOTE.

Di sisi lain, (Javeed et al., 2023) merancang model prediksi demensia dengan mengombinasikan Feature Extraction Battery (FEB) dan SVM. Penelitian ini berangkat dari keterbatasan model machine learning sebelumnya yang memiliki akurasi rendah serta rentan terhadap bias. Dataset yang digunakan berasal dari Swedish National Study on Aging and Care (SNAC) yang mencakup parameter fisik, psikologis, sosial, serta gaya hidup responden. Hasil penelitian menunjukkan bahwa model FEB-SVM unggul dibandingkan 12 model machine learning pembanding. Pada data uji, metode ini memperoleh akurasi 93,92%, lebih tinggi sekitar 6% dibandingkan SVM konvensional. Selain itu, diperoleh nilai presisi 91,80%, recall 86,59%, F1-score 89,12%, serta Matthew's Correlation Coefficient (MCC)

sebesar 0,4987. Keunggulan performa ini terutama diperoleh karena FEB berhasil mereduksi dimensi dari 75 fitur menjadi hanya 9 fitur penting, sehingga kompleksitas model menurun dan risiko overfitting dapat diminimalisasi.

Pencegahan dan pengendalian Diabetes Melitus kerap dilakukan dengan menerapkan pola hidup sehat, dalam hal ini agar orang yang memiliki faktor risiko dan penderita Diabetes Melitus dapat mengendalikan penyakitnya sehingga tidak terjadi komplikasi atau kematian dini (American Diabetes Association Professional Practice Committee, 2022). Upaya pencegahan dan pengendalian ini dilakukan melalui edukasi dan tatalaksana sesuai standar. Deteksi dini Diabetes Melitus sangat diperlukan, dengan demikian penelitian ini dilakukan untuk klasifikasi deteksi awal risiko Diabetes Melitus dengan memanfaatkan penggunaan Machine Learning.(Junus et al., 2023)

Split Dataset adalah proses membagi dataset menjadi beberapa bagian untuk tujuan pelatihan (*training*) dan pengujian (*testing*) model machine learning Pendekatan ini penting terutama pada dataset yang memiliki imbalance class, seperti pada kasus deteksi penyakit, di mana jumlah data pasien positif (class 1) biasanya lebih sedikit dibandingkan data negatif (class 0). Dengan proporsi yang seimbang, model pembelajaran mesin akan mendapatkan representasi yang lebih baik dari kedua kelas, sehingga hasil evaluasi kinerja model dapat lebih mencerminkan kemampuan generalisasi terhadap data baru (Chawla et al., 2002)

Salah satu pendekatan yang berkembang pesat dalam bidang diagnosis prediktif adalah penggunaan algoritma pembelajaran mesin (*machine learning*). Salah satu algoritma yang terbukti unggul dalam klasifikasi adalah *Support Vector Machine* (SVM), terutama ketika dikombinasikan dengan *Kernel Radial Basis Function* (RBF). SVM-RBF memiliki keunggulan dalam menangani data non-linear an kompleks yang sering ditemui dalam dataset kesehatan (Koklu & Sulak, 2024).

Tantangan yang sering muncul dalam penerapan algoritma SVM pada data medis adalah ketidakseimbangan kelas (*imbalanced dataset*), di mana jumlah data penderita diabetes biasanya lebih sedikit dibandingkan data non-diabetes. Ketidakseimbangan ini dapat menyebabkan bias dalam model prediksi. Oleh karena itu, teknik *oversampling* seperti *Adaptive Synthetic Sampling* (ADASYN) digunakan untuk menyeimbangkan distribusi kelas dalam dataset (Haibo He et al., 2008). Selain itu, normalisasi data menggunakan StandarScaler juga penting untuk memastikan setiap fitur memiliki skala yang sama, sehingga meningkatkan efisiensi pembelajaran algoritma (Han & Kamber, 2012).

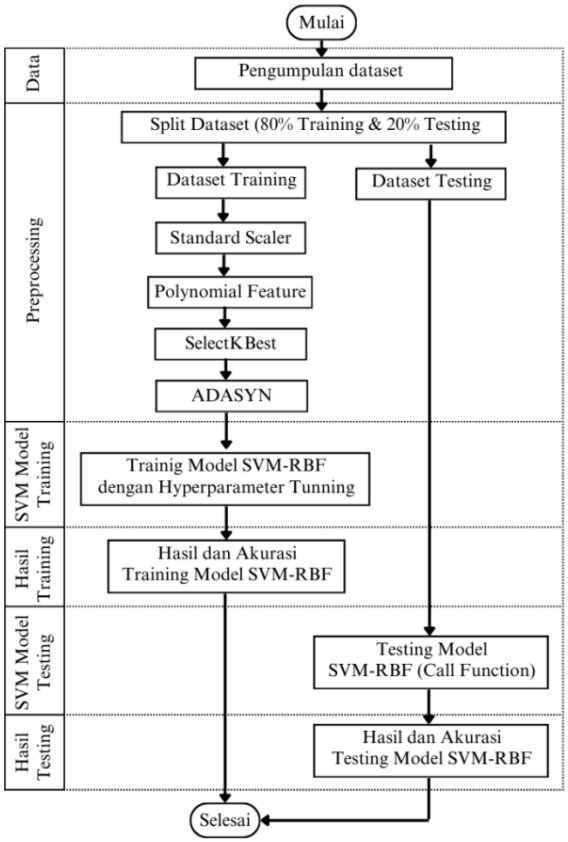
Dengan mempertimbangkan tantangan tersebut, penelitian ini bertujuan untuk mengevaluasi kinerja algoritma SVM dengan kernel RBF dalam mengklasifikasikan diabetes, serta mengkaji pengaruh *preprocessing data* berupa ADASYN dan Untuk memperoleh hasil analisis yang akurat, perlu dilakukan proses standarisasi dan penyesuaian skala pada fiturfitur yang dipilih sebelum dilakukan prognosis dan pembuatan model. Untuk melakukan proses tersebut, digunakan fungsi StandardScaler dari library Scikit-learn (Haibo He et al., 2008) StandartScaler adalah metode pra-pemrosesan data yang sering digunakan dalam machine learning untuk menormalkan fitur. Tujuannya adalah memastikan setiap fitur memiliki distribusi dengan rata-rata 0 (mean = 0) dan simpangan baku 1 (std = 1).

Penggunaan *PolynomialFeatures degree = 2* memungkinkan model mengakses fitur non-linear seperti kuadrat (misalnya Glucose², BMI²) dan interaksi antar fitur (Glucose × BMI, Age × Pregnancies), yang memperkaya representasi data dan meningkatkan kemampuan SVM dalam menangkap pola non-linear tersembunyi (Ejiyi et al., 2025). Pada tahap preprocessing, dipilih fitur teratas menggunakan SelectKBest dengan fungsi skor *mutual information* (mutual_info_classif), karena metode ini dapat menangkap hubungan non-linier antara fitur dan target Outcome—membantu mengurangi jumlah fitur dan

mencegah overfitting setelah ekspansi fitur via PolynomialFeatures (k=20) digunakan untuk memilih fitur paling informatif (Bhagya, 2025)

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tujuan menguji performa *algoritma Support Vector Machine* (SVM) dengan *kernel Radial Basis Function* (RBF) dalam klasifikasi penyakit Diabetes Mellitus. Fokus utama penelitian ini adalah mengukur peningkatan performa model melalui tahapan preprocessing data, termasuk normalisasi, transformasi fitur non-linear, seleksi fitur, dan penyeimbangan kelas. Adapun tahapan metodologi penelitian dijelaskan sebagai berikut.



Gambar 1. Alur Penelitian

2.1.Pengumpulan Dataset

Tahap awal penelitian adalah pengumpulan dataset yang menjadi dasar bagi pembangunan model prediksi. Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang diperoleh dari Kaggle. Dataset ini berisi 768 sampel data pasien dengan berbagai variabel kesehatan sebagai fitur prediktor, serta label target berupa status diabetes.

2.2.Pembagian Dataset

Setelah dataset terkumpul, langkah selanjutnya adalah melakukan pembagian data menjadi data training dan data testing. Pada penelitian ini, data dibagi dengan perbandingan 80% untuk data training dan 20% untuk data testing. Pembagian ini dilakukan agar model dapat dilatih menggunakan sebagian besar data, sementara evaluasi dilakukan pada data yang belum pernah dilihat sebelumnya untuk menguji kemampuan generalisasi model.

2.3.Preprocessing Data

Preprocessing data dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam training model. Proses ini hanya diterapkan pada data training agar tidak menimbulkan data leakage, dan hasil transformasi kemudian diaplikasikan juga ke data testing. Tahapan preprocessing yang dilakukan adalah sebagai berikut:

- Standard Scaler: proses normalisasi dilakukan untuk memastikan setiap fitur memiliki distribusi dengan rata-rata nol dan standar deviasi satu, sehingga mengurangi pengaruh perbedaan skala antar variabel.
- Polynomial Feature: digunakan untuk membangkitkan kombinasi fitur nonlinier, sehingga hubungan kompleks antar variabel dapat ditangkap dengan lebih baik oleh algoritma SVM.
- SelectKBest: metode seleksi fitur digunakan untuk memilih variabel yang paling berpengaruh terhadap target, dengan tujuan mengurangi dimensi dan meningkatkan efisiensi training.

- ADASYN (Adaptive Synthetic Sampling): teknik oversampling diterapkan untuk mengatasi ketidakseimbangan kelas, dengan cara menghasilkan sampel sintetis adaptif pada kelas minoritas agar distribusi data menjadi lebih seimbang.

2.4.Training model

Tahap training dilakukan dengan menggunakan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF). Model SVM dipilih karena kemampuannya dalam menangani data berdimensi tinggi dan relasi nonlinier antar fitur. Untuk meningkatkan performa model, dilakukan tuning hyperparameter melalui eksplorasi nilai parameter berikut:

- C: parameter regulasi yang mengontrol keseimbangan antara margin maksimum dan tingkat kesalahan klasifikasi.
- Gamma: parameter kernel RBF yang menentukan seberapa jauh pengaruh sebuah sampel data terhadap pembentukan boundary.
- Degree: digunakan sebagai eksplorasi tambahan ketika fitur polinomial diaplikasikan agar model mampu mempelajari interaksi nonlinier yang lebih kompleks.

Pencarian nilai terbaik dilakukan dengan distribusi log-uniform agar rentang pencarian lebih luas dan efisien.

2.5.Evaluasi Model pada Data Training

Setelah proses training selesai, model dievaluasi terlebih dahulu menggunakan data training untuk mengukur performa awal. Evaluasi ini bertujuan untuk mengetahui apakah model mampu mempelajari pola dari data dengan baik. Metrik yang digunakan pada tahap ini meliputi akurasi, presisi, recall, dan F1-score.

2.6.Testing Model

Langkah berikutnya adalah menerapkan model terbaik hasil tuning hyperparameter pada data testing. Testing ini dilakukan untuk mengukur sejauh mana model dapat melakukan generalisasi terhadap data baru yang belum pernah dipelajari sebelumnya. Dengan demikian, tahap ini menjadi indikator utama keberhasilan model dalam menyelesaikan permasalahan klasifikasi diabetes.

2.7.Evaluasi Model pada Data testing

Tahap terakhir adalah evaluasi kinerja model berdasarkan hasil prediksi pada data testing. Evaluasi dilakukan menggunakan metrik klasifikasi seperti akurasi, presisi, recall, dan F1-score untuk mendapatkan gambaran menyeluruh terkait keunggulan maupun kelemahan model.

3. HASIL 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.kemampuan model dalam mendeteksi kasus diabetes (kelas minoritas).

3.1. Pengumpulan Dataset

Data yang digunakan dalam penelitian ini bersumber dari platform Kaggle, yaitu dataset yang sangat dikenal bernama *Pima Indians Diabetes Dataset*. Dataset ini banyak digunakan dalam penelitian klasifikasi diabetes, terutama karena memiliki data medis nyata dan tantangan tersendiri akibat ukuran data yang terbatas serta distribusi kelas yang tidak seimbang.

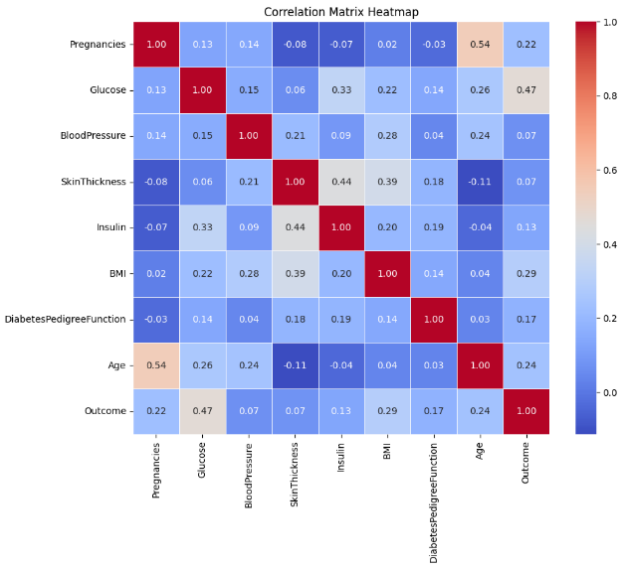
Dataset ini terdiri atas 768 data pasien perempuan suku Indian Pima di Amerika

Serikat. Dataset memuat total 9 kolom fitur yang meliputi berbagai informasi medis, seperti ditunjukkan pada Tabel 1.

Tabel 1. Fitur Dataset

No	Nama Kolom	Deskripsi
1	Pregnancies	Jumlah kehamilan
2	Glucose	Kadar glukosa plasma saat tes
3	BloodPressure	Tekanan darah diastolik (mm Hg)
4	SkinThickness	Ketebalan lipatan kulit trisep (mm)
5	Insulin	Kadar insulin serum (mu U/ml)
6	BMI	Indeks massa tubuh (kg/m²)
7	DPF	Fungsi silsilah diabetes
8	Age	Usia (tahun)
9	Outcome	Target klasifikasi:

Ketidakseimbangan jumlah data pada kelas target, di mana data pasien yang tidak menderita diabetes (kelas 0) berjumlah 500 atau sekitar 65,1%, sedangkan data pasien yang menderita diabetes (kelas 1) hanya sebanyak 268 atau sekitar 34,9%, dapat menyebabkan model cenderung lebih akurat dalam memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Oleh karena itu, diperlukan teknik penyeimbangan data seperti ADASYN, yang bertujuan menambah data sintetis pada kelas minoritas agar distribusi menjadi lebih seimbang dan model dapat belajar dengan baik dari kedua kelas. Selain itu, dataset ini memuat kolom numerik dengan rentang nilai yang sangat bervariasi, contohnya kolom Glucose dengan nilai hingga 199, kolom BMI yang berkisar sekitar 23,3–43,1, serta DiabetesPedigreeFunction yang hanya bernilai kecil seperti 0,167 hingga 2,288. Perbedaan skala yang cukup besar ini dapat membuat algoritma SVM bias terhadap fitur dengan nilai lebih tinggi. Untuk mengatasi masalah ini, dilakukan penyeragaman data menggunakan teknik normalisasi melalui *StandardScaler*, sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1, dan model dapat memproses semua fitur secara seimbang tanpa dipengaruhi skala aslinya.



Gambar 2. Matriks Korelasi Fitur terhadap Target

Analisis awal terhadap hubungan antar fitur dilakukan menggunakan matriks korelasi, seperti ditunjukkan pada Gambar 1. Hasil matriks korelasi pada gambar di atas menunjukkan bahwa sebagian besar fitur memiliki nilai korelasi yang relatif rendah terhadap target Outcome. Sebagai contoh, Glucose memiliki korelasi tertinggi sekitar 0,47, disusul oleh BMI (0,29), Age (0,24), dan Pregnancies (0,22). Sementara fitur-fitur lain seperti BloodPressure, SkinThickness, dan Insulin hanya memiliki korelasi sangat rendah di bawah 0,15. Korelasi antar fitur pun cenderung rendah hingga sedang, misalnya antara SkinThickness dengan Insulin (0,44) atau SkinThickness dengan BMI (0,39).

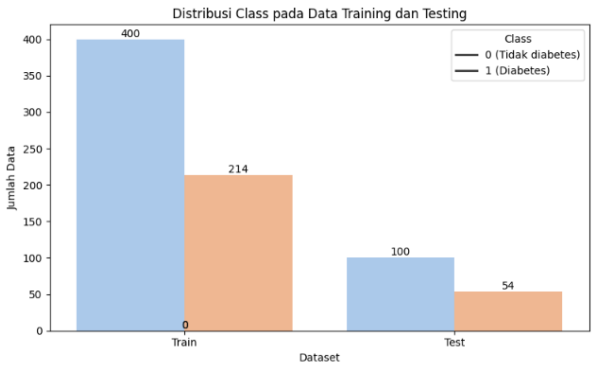
Berdasarkan temuan ini, dibuatlah rangkaian tahapan preprocessing dalam code, seperti Polynomial Features degree=2 untuk

membangkitkan fitur interaksi non-linear antar variabel yang awalnya hanya berkorelasi rendah, dan SelectKBest untuk memilih 20 fitur terbaik yang paling informatif terhadap target. Teknik ini penting karena meskipun fitur aslinya memiliki korelasi rendah, model SVM dapat tetap menangkap pola non-linear yang muncul setelah transformasi polinomial.

3.2. Split Dataset

Pada tahap ini, dataset dibagi menjadi dua bagian menggunakan fungsi train_test_split dengan proporsi 80% untuk data latih (training set) dan 20% untuk data uji (testing set). Pembagian ini tidak dilakukan secara acak penuh, melainkan menggunakan parameter stratify. Tujuannya adalah agar distribusi kelas target (Outcome) tetap proporsional antara data latih dan data uji (ditunjukkan pada Gambar 2). Dengan cara ini, model diharapkan dapat belajar dari data yang mewakili kondisi sebenarnya, sekaligus mengurangi risiko bias akibat ketidakseimbangan kelas yang terjadi pada salah satu subset.

Hasil dari proses pembagian ini adalah data latih berisi total 614 data, yang terdiri atas class 0 (tidak menderita diabetes) sebanyak 400 data, dan class 1 (menderita diabetes) sebanyak 214 data. Sedangkan pada data uji yang berjumlah 154 data, terdapat 100 data class 0 dan 54 data class 1. Pembagian yang seimbang seperti ini penting untuk memastikan bahwa evaluasi performa model nantinya benar-benar mencerminkan kemampuan generalisasi model terhadap data baru, baik untuk kelas mayoritas maupun kelas minoritas.



Gambar 3. Distribusi Class pada Data Training dan Testing

3.3. Tahapan Preprocessing

Preprocessing merupakan sebuah langkah penting dalam proses penambangan data (Muhammad Hilmy Haidar Aly, 2024).Tahapan preprocessing diterapkan untuk meningkatkan kualitas data yang masuk ke model dan membantu algoritma SVM bekerja lebih baik, terutama karena SVM sensitif terhadap skala dan distribusi data. Adapun preprocessing yang dilakukan meliputi:

3.3.1. StandardScaler

Pada tahap StandardScaler, dilakukan normalisasi terhadap semua fitur numerik agar memiliki rata-rata (*mean*) 0 dan standar deviasi 1. Tujuan normalisasi ini adalah untuk menyamakan skala antar fitur yang awalnya memiliki rentang nilai sangat berbeda, seperti kolom Glucose yang dapat mencapai nilai maksimum 199, sedangkan Diabetes Pedigree Function hanya sekitar 2,42. Tanpa normalisasi, algoritma SVM dapat secara tidak sengaja memberi bobot lebih besar pada fitur dengan skala lebih tinggi, sehingga model menjadi bias dan sulit mempelajari hubungan antar fitur secara adil. Normalisasi juga penting untuk menjaga stabilitas perhitungan jarak antar data, terutama pada SVM berbasis kernel seperti RBF yang sensitif terhadap perbedaan skala.

Penerapan StandardScaler sebelum tahap feature engineering seperti PolynomialFeatures degree=2 menjadi langkah yang tepat. Transformasi polynomial akan menghasilkan fitur baru berupa kuadrat dan interaksi antar fitur; jika dilakukan tanpa normalisasi terlebih dahulu, fitur-fitur hasil transformasi bisa memiliki nilai yang sangat besar dan mendominasi model. Dengan menormalisasi terlebih dahulu, fitur asli sudah berada pada skala yang seragam sehingga fitur polinomial yang terbentuk juga tetap terkontrol. Pendekatan ini membantu model SVM mempelajari pola data secara lebih seimbang dan meningkatkan kemampuan

generalisasi, seperti tercermin dari hasil evaluasi model yang stabil pada data uji.

3.3.2. PolynomialFeatures (degree=2)

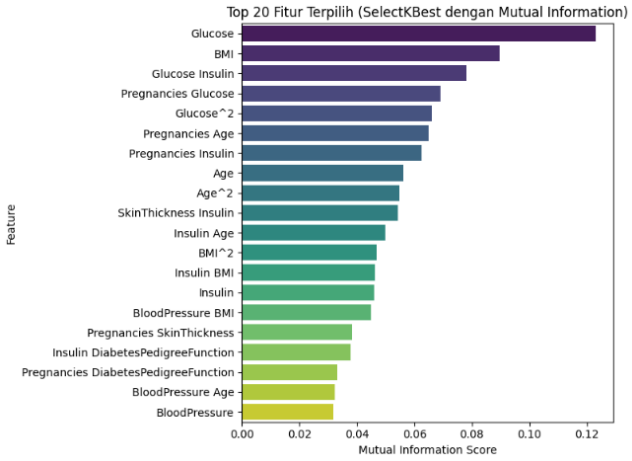
Pada tahap PolynomialFeatures degree=2, dilakukan transformasi data untuk menghasilkan fitur-fitur baru berupa kuadrat dan interaksi antar fitur asli. Proses ini bertujuan memperkaya representasi data sehingga model dapat menangkap pola-pola non-linear yang tidak dapat dijelaskan hanya dengan fitur asli. Dalam konteks dataset diabetes ini, beberapa fitur seperti Glucose atau BMI memang memiliki korelasi sedang terhadap target Outcome, namun fitur lainnya menunjukkan korelasi yang relatif rendah. Dengan menambahkan fitur hasil perkalian antar variabel (seperti Glucose × BMI, Age × Pregnancies, dan sebagainya) maupun kuadrat masing-masing variabel (seperti $Glucose^2$ atau BMI^2), model SVM diharapkan dapat mendeteksi hubungan non-linear yang tersembunyi dan lebih kompleks.

PenggunaanPolynomialFeatures degree=2 menjadi penting karena algoritma SVM sendiri memiliki keunggulan dalam mempelajari pola non-linear, terutama saat dipadukan dengan kernel seperti RBF. Transformasi polynomial ini menghasilkan total 44 fitur baru yang terdiri dari fitur asli, kuadrat setiap fitur, serta interaksi dua-dua antar fitur.

Meski jumlah fitur meningkat, potensi overfitting dapat dikurangi melalui tahap selanjutnya, yaitu SelectKBest, yang akan memilih fitur paling relevan. Dengan strategi ini, model tidak hanya bergantung pada hubungan linear, tetapi juga mampu mempertimbangkan efek gabungan antar fitur yang mungkin lebih signifikan dalam menentukan risiko diabetes pada pasien.

3.3.3. SelectKBest

Pada tahap preprocessing, diterapkan SelectKBest dengan fungsi skor mutual information (mutual_info_classif) untuk memilih sejumlah fitur yang paling relevan terhadap target Outcome. Secara teknis, SelectKBest bekerja dengan menghitung skor relevansi setiap fitur terhadap target menggunakan metode statistic mutual information, yang mampu menangkap tidak hanya hubungan linear tetapi juga non-linear antara masing-masing fitur dengan label target. Proses ini penting, terutama setelah tahap PolynomialFeatures degree=2 yang menghasilkan total 44 fitur (fitur asli, kuadrat, dan interaksi). Dari semua fitur tersebut, tidak semuanya informatif atau relevan, sehingga perlu dilakukan seleksi agar model tidak belajar dari noise atau informasi yang tidak signifikan. Dalam penelitian ini, parameter k=20 ditentukan untuk memilih 20 fitur dengan skor mutual information tertinggi. Nilai k=20 ini diujikan agar tetap memanfaatkan sebagian besar informasi dari fitur polinomial tanpa menyebabkan model menjadi terlalu kompleks dan rawan overfitting, sebagaimana hasil seleksi fitur yang ditunjukkan pada Gambar 3.



Gambar 4. Top 14 Fitur Terpilih Berdasar Skor Mutual Information

3.3.4. ADASYN (Adaptive Synthetic Sampling)

Pada tahap preprocessing, salah satu langkah penting yang diterapkan adalah penggunaan teknik oversampling adaptif ADASYN (*Adaptive Synthetic Sampling*). Teknik ini dirancang khusus untuk menangani

masalah imbalanced dataset, yaitu ketidakseimbangan distribusi kelas target antara pasien yang menderita diabetes dan yang tidak. Berdasarkan distribusi awal dataset, jumlah data untuk kelas 0 (tidak menderita diabetes) sebanyak 500 data (65,1%), sedangkan kelas 1 (menderita diabetes) hanya 268 data (34,9%). Ketidakseimbangan seperti ini dapat menyebabkan model cenderung "berpihak" pada kelas mayoritas, sehingga mengurangi

ADASYN bekerja dengan menghasilkan data synthetic pada area data minoritas yang sulit diklasifikasikan yakni data minoritas yang berada di sekitar batas keputusan (decision boundary). ADASYN bersifat adaptif: jumlah synthetic samples yang ditambahkan disesuaikan dengan tingkat kesulitan lokal dari data tersebut. Pada penelitian ini, ADASYN diterapkan di dalam pipeline preprocessing setelah tahap StandardScaler, PolynomialFeatures, dan SelectKBest. Pemrosesan dilakukan hanya pada data latih dalam setiap fold cross-validation agar mencegah data leakage dan menjaga validitas model.

Hasil balancing dengan parameter sampling_strategy = 0.75 menunjukkan bahwa jumlah data minoritas pada data latih bertambah dari semula 214 menjadi 314 setelah diterapkan ADASYN, sementara jumlah data mayoritas tetap sebanyak 400. Artinya, ADASYN secara otomatis memutuskan untuk menambahkan sekitar 100 data synthetic agar rasio kelas minoritas mendekati target (minority ≈ 75% dari majority). Pendekatan ini bukan hanya meningkatkan jumlah data minoritas, tetapi juga memperkaya representasi data minoritas terutama di area yang sulit diprediksi, sehingga membantu model SVM RBF dalam menangkap pola non-linear yang lebih kompleks dan meningkatkan kemampuan generalisasi model terhadap kasus diabetes pada data uji.

3.4 Training, Testing dan Evaluasi Model

Proses training model dilakukan menggunakan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) yang dirancang khusus untuk mengenali pola non-linear. Sebelum tahap training, data telah diproses melalui pipeline yang meliputi normalisasi fitur dengan StandardScaler, pembentukan fitur non-linear menggunakan PolynomialFeatures degree=2, pemilihan 20 fitur terbaik melalui SelectKBest dengan metode mutual information, serta penyeimbangan data menggunakan ADASYN. Pipeline ini dirancang untuk memperkuat representasi data dan menangani masalah ketidakseimbangan kelas dalam dataset.

Data latih (training set) yang berjumlah 614 data (400 untuk kelas tidak diabetes dan 214 untuk kelas diabetes sebelum ADASYN) kemudian dibagi menjadi masukan dalam proses training, sebagaimana ditunjukkan pada Tabel 2. Kemudian menjadi masukan dalam proses training. Model dilatih agar dapat mengenali pola yang memisahkan kedua kelas sebaik mungkin, dengan memanfaatkan fitur-fitur asli dan hasil transformasi polinomial. Setelah training selesai, model diuji pada data uji (testing set) sebanyak 154 data, yang belum pernah dilihat model sebelumnya. Pengujian ini penting untuk menilai kemampuan generalisasi model terhadap data baru.

Tabel 2. Distribusi Data Training dan Testing Sebelum Oversampling ADASYN

Class	Training Set (Asli)	Testing Set
0	400	100
1	214	54
Total	**614**	**154**

Dalam tahap hyperparameter tuning, model dieksplorasi dengan ruang parameter cukup luas, seperti nilai C, gamma, dan degree polinomial,menggunakan Randomized Search CV. Hasil tuning menunjukkan parameter terbaik (poly_degree=2, svc_C=60.15, svc_gamma=0.000826) seperti yang ditunjukkan pada Tabel 3. Konfigurasi ini memungkinkan model lebih adaptif terhadap

data yang bersifat non-linear dan mengatasi tantangan ketidakseimbangan kelas dengan memberikan margin optimal antara data positif dan negatif.

Tabel 3. Hasil Tuning Hyperparameter

Parameter	Nilai Terbaik
poly__degree	2
svc__C	60.15
svc__gamma	0.000826

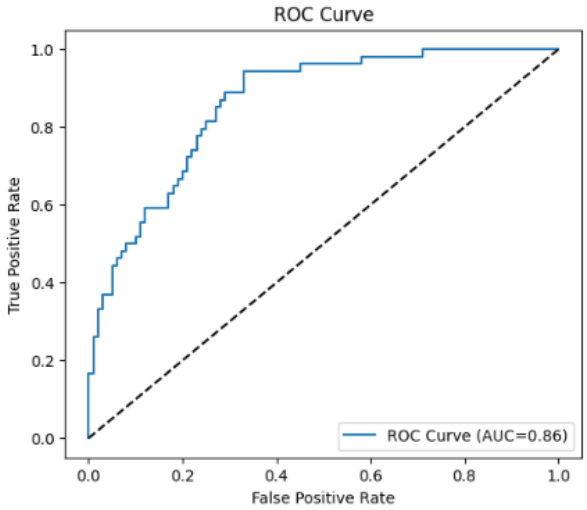
Hasil evaluasi pada data training menunjukkan performa yang cukup baik dengan Accuracy sebesar 0.779, Precision 0.81, Recall 0.779, dan F1-Score 0.783. Angka-angka ini menunjukkan bahwa model tidak hanya mampu mengklasifikasikan data mayoritas dengan baik, tetapi juga relatif sensitif dalam mendeteksi data minoritas (penderita diabetes). Hasil evaluasi model SVM RBF pada data training dan testing dapat dilihat pada Tabel 4. Tabel ini menampilkan nilai Accuracy, Precision, Recall, F1-Score, dan ROC AUC yang digunakan untuk mengukur performa model dalam membedakan pasien yang menderita diabetes dan yang tidak, meskipun dataset aslinya imbalanced.

Tabel 4. Hasil Evaluasi Model SVM RBF

Metric	Data Training	Data Testing
Accuracy	0.779	0.766
Precision	0.81	0.792
Recall	0.779	0.766
F1-Score	0.783	0.771
ROC AUC	–	0.861

Keberhasilan model dalam menghasilkan nilai ROC AUC di atas 0.86 dan F1-Score mendekati 0.77 dapat dijelaskan oleh kombinasi preprocessing dan pemilihan hyperparameter. Teknik PolynomialFeatures degree=2 membantu menangkap interaksi non-linear antar fitur, sementara SelectKBest memastikan hanya fitur yang paling informatif yang digunakan. ADASYN juga berperan penting dalam memperkaya data minoritas, sehingga model tidak hanya belajar dari pola mayoritas, tetapi juga mampu mendeteksi pasien yang sebenarnya berisiko namun jumlahnya lebih sedikit.

Selain itu, kinerja model juga divisualisasikan menggunakan kurva *Receiver Operating Characteristic* (ROC) untuk menilai kemampuan model dalam membedakan kelas positif dan negatif, seperti ditunjukkan pada Gambar 4.



Gambar 5. Kurva ROC Model

3.5 Pembahasan

Model SVM dengan kernel RBF yang dikembangkan dalam penelitian ini menunjukkan performa yang cukup baik berdasarkan hasil evaluasi di data training maupun testing. Pada data training, model berhasil mencapai Accuracy sebesar 0.779, Precision 0.81, Recall 0.779, dan F1-Score 0.783. Nilai-nilai ini menandakan bahwa model tidak hanya mampu mengenali pola yang ada dalam data mayoritas, tetapi juga cukup sensitif dalam mendeteksi kelas minoritas, yaitu pasien yang menderita diabetes. Sementara itu, evaluasi pada data testing—yang menjadi ukuran sejauh mana model dapat menggeneralisasi ke data baru—menunjukkan Accuracy sebesar 0.766, Precision 0.792, Recall 0.766, F1-Score 0.771, serta ROC AUC sebesar 0.861. Nilai ROC AUC yang cukup tinggi ini mengindikasikan bahwa model mampu membedakan dengan baik antara pasien diabetes dan non-diabetes,

meskipun dataset aslinya memiliki ketidakseimbangan kelas.

Jika dibandingkan dengan penelitian terdahulu (Kumar, 2022) yang menggunakan dataset serupa, seperti model SVM RBF yang dikembangkan peneliti sebelumnya hanya memperoleh akurasi sebesar 75.52%. Nilai tersebut sedikit lebih rendah dibandingkan hasil akurasi model dalam penelitian ini, yang mencapai 76.6%. Perbedaan ini dapat diinterpretasikan sebagai hasil dari eksplorasi hyperparameter yang lebih optimal dan pendekatan pipeline yang lebih sistematis dalam penelitian ini, meskipun secara keseluruhan selisihnya tidak terlalu besar. Hal ini memperlihatkan bahwa dengan tuning parameter dan pemilihan fitur yang tepat, kinerja model SVM RBF dapat ditingkatkan walau hanya beberapa persen tetapi cukup signifikan untuk meningkatkan kemampuan deteksi risiko diabetes. Untuk melihat posisi performa penelitian ini dibandingkan penelitian sebelumnya, dilakukan perbandingan akurasi dan konteks pengujian seperti ditunjukkan pada Tabel 5.

Tabel 5. Perbandingan Hasil Penelitian

Penelitian	Algoritma	Dataset	Accuracy	Keterangan
Penelitian ini (2025)	SVM RBF	Pima Indians Diabetes	**76.60%**	Dengan pipeline lengkap: StandardScaler, PolynomialFeatures,SelectKBest, ADASYN, tuning hyperparameter
Penelitian lain	SVM RBF	Pima Indians Diabetes	75.52%	Tanpa dijelaskan detail preprocessing dan tuning yang sama mendalam

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa pendekatan pipeline yang terstruktur, penggunaan transformasi non-linear, balancing data, serta eksplorasi hyperparameter yang luas dapat meningkatkan performa model SVM RBF. Model yang dihasilkan tidak hanya baik di data latih, tetapi juga mampu mempertahankan performa cukup stabil di data uji, sehingga lebih dapat diandalkan dalam konteks prediksi risiko diabetes berdasarkan dataset Pima. Pendekatan ini membuktikan bahwa meskipun algoritma SVM sangat bergantung pada kualitas fitur dan distribusi data, dengan strategi preprocessing dan tuning yang tepat, akurasi dan generalisasi model dapat ditingkatkan di atas baseline penelitian terdahulu.

4. PENUTUP

4.1. Kesimpulan

Penelitian ini berhasil membuktikan bahwa algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) dapat memberikan performa yang baik dalam klasifikasi penyakit Diabetes Mellitus, terutama ketika dikombinasikan dengan teknik preprocessing yang tepat. Penggunaan pipeline preprocessing yang meliputi normalisasi dengan StandardScaler, pembangkitan fitur non-linear dengan PolynomialFeatures degree=2, seleksi fitur melalui SelectKBest, serta penyeimbangan data menggunakan ADASYN mampu meningkatkan kemampuan model dalam mendeteksi kelas minoritas (penderita diabetes) yang sebelumnya sulit dikenali. Evaluasi model menunjukkan nilai ROC AUC sebesar 0.861 dan F1-Score sebesar 0.771 pada data testing, yang mencerminkan kemampuan model dalam menggeneralisasi dan membedakan kelas secara efektif, meskipun pada dataset yang tidak seimbang.

Dengan demikian, pendekatan berbasis machine learning ini berpotensi menjadi alat bantu yang andal untuk deteksi dini diabetes, terutama dalam kondisi di mana akses terhadap tenaga medis terbatas dan diagnosis manual berisiko terlewatkan..

4.2. Saran

Untuk pengembangan lebih lanjut, terdapat beberapa saran yang dapat dipertimbangkan:
Diversifikasi Dataset: Penggunaan dataset yang lebih beragam, baik dari segi populasi geografis maupun jenis kelamin, dapat meningkatkan generalisasi model dalam praktik nyata.
Pengujian Model di Data Nyata (Real-World Deployment): Implementasi model ini dalam sistem klinis atau aplikasi kesehatan digital akan memberikan gambaran lebih konkret

tentang efektivitas dan kegunaannya di lapangan.

Eksplorasi Algoritma Lain: Penelitian selanjutnya dapat membandingkan performa SVM-RBF dengan algoritma lain seperti XGBoost, Random Forest, atau deep learning untuk mengidentifikasi pendekatan terbaik.
Penggabungan dengan Fitur Tambahan: Menambahkan data eksternal seperti riwayat keluarga, pola makan, dan aktivitas fisik dapat memperkaya fitur dan meningkatkan akurasi prediksi.
Dengan menerapkan saran-saran tersebut, diharapkan penelitian klasifikasi diabetes menggunakan pendekatan machine learning akan semakin relevan dan berdampak nyata dalam mendukung sistem kesehatan masyarakat.

5. DAFTAR PUSTAKA

American Diabetes Association Professional Practice Committee. (2022). 9. Pharmacologic Approaches to Glycemic Treatment: *Standards of Medical Care in Diabetes—2022*. *Diabetes Care*, *45*(Supplement_1), S125–S143. https://doi.org/10.2337/dc22-S009

Bhagya, R. (2025). 10 Feature Selection Techniques Built into Scikit-learn That Every Data Scientist Should Know. *Medium*. https://medium.com/%40bhagyarana80/10-feature-selection-techniques-built-into-scikit-learn-that-every-data-scientist-should-know-f63bc5fb77d7

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

Ejiyi, C. J., Cai, D., Eze, F. O., Ejiyi, M. B., Idoko, J. E., Asere, S. K., & Ejiyi, T. U. (2025). Polynomial-SHAP as a SMOTE alternative in conglomerate neural networks for realistic data augmentation in cardiovascular and breast cancer diagnosis. *Journal of Big Data*, *12*(1), 97. https://doi.org/10.1186/s40537-025-01152-3

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

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

IDF. (2019). *The Diabetes Atlas*. https://diabetesatlas.org/

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), 439. 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

Kemenkes RI. (2021). *PROFIL KESEHATAN INDONESIA 2020*. Kementerian Kesehatan RI.

Koklu, N., & Sulak, S. A. (2024). Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. *Sinop Üniversitesi Fen Bilimleri Dergisi*, *9*(1), 217–239.

https://doi.org/10.33484/sinopfbd.1445215

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

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

Ruze, R., Liu, T., Zou, X., Song, J., Chen, Y., Xu, R., Yin, X., & Xu, Q. (2023). Obesity and type 2 diabetes mellitus: Connections in epidemiology, pathogenesis, and treatments. *Frontiers in Endocrinology*, *14*, 1161521. https://doi.org/10.3389/fendo.2023.1161521

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Transactions on Computer-Human Interaction*, *27*(5), 1–53. https://doi.org/10.1145/3398069

Wong, J. E., Yamaguchi, M., Nishi, N., Araki, M., & Lee, L. H. (n.d.). *Predicting overweight and obesity status among Malaysian working adults: Comparing performance of machine learning with logistic regression*.

World Health Organization, Diabetes. (2021). WHO. https://www.who.int/news-room/fact-sheets/detail/diabetes