PENERAPAN *SUPPORT VECTOR MACHINE* DENGAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) DALAM PENGKLASIFIKASIAN PENYAKIT DIABETES

SY. FARINI NURHALIZA

NIM H1091211037

SKRIPSI

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS TANJUNGPURA

PONTIANAK

2025

PENERAPAN *SUPPORT VECTOR MACHINE* DENGAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) DALAM PENGKLASIFIKASIAN PENYAKIT DIABETES

SY. FARINI NURHALIZA

NIM H1091211037

SKRIPSI

Sebagai salah satu syarat untuk memperoleh gelar
Sarjana Statistika pada Program Studi Statistika

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

LEMBAR PENGESAHAN

Judul Tugas Akhir	: Penerapan *Support Vector Machine* dengan *Synthetic Oversampling Minority Technique (SMOTE)* dalam Pengklasifikasian Penyakit Diabetes
Nama Mahasiswa	: Sy. Farini Nurhaliza
NIM	: H1091211037
Jurusan/Program Studi	: Matematika/Statistika
Tanggal Lulus	: 6 Mei 2025
SK Pembimbing	: No. 2616/UN22.8/TD.06/2024/Tanggal 2 September 2024
SK Penguji	: No. 1271/UN22.8/TD.06/2025/Tanggal 28 April 2025

Dosen Pembimbing

Pembimbing I	Pembimbing II
Nurfitri Imro'ah, M.Si.	Shantika Martha, M.Si.
NIP 198907182019032021	NIP 198403082008122003

Dosen Penguji

Ketua Penguji	Anggota Penguji
Dr. Evy Sulistianingsih, M.Sc.	Hendra Perdana, M.Sc.
NIP 198502172008122006	NIP 198810102019031020
Pimpinan Sidang (merangkap anggota penguji)	Sekretaris Sidang (merangkap anggota penguji)
Nurfitri Imro'ah, M.Si.	Shantika Martha, M.Si.
NIP 198907182019032021	NIP 198403082008122003

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Dr. Gusrizal, M.Si.
NIP 197108022000031001

PERNYATAAN INTEGRITAS AKADEMIK

Saya yang bertanda tangan di bawah ini:

Nama : Sy. Farini Nurhaliza

NIM : H1091211037

Program Studi/ Jurusan : Statistika/Matematika

Fakultas : Matematika dan Ilmu Pengetahuan Alam

dengan ini menyatakan bahwa dokumen ilmiah Tugas Akhir yang disajikan ini tidak mengandung unsur pelanggaran integritas akademik sesuai Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia Nomor 39 Tahun 2021. Apabila di kemudian hari dokumen ilmiah Tugas Akhir ini mengandung unsur pelanggaran integritas akademik sesuai ketentuan perundangan tersebut, maka saya bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Demikian pernyatan ini untuk digunakan sebagaimana mestinya.

Pontianak, 6 Mei 2025

Sy. Farini Nurhaliza
NIM H1091211037

Penerapan *Support Vector Machine* Dengan *Synthetic Oversampling Minority Technique (SMOTE) Dalam Pengklasifikasian* Pasien Diabetes

Abstrak

Algoritma *Support Vector Machine* (SVM) merupakan salah satu *algoritma* pada *data mining* yang sering digunakan dalam proses klasifikasi karena kemampuannya menangani data dengan dimensi tinggi. Namun, performa klasifikasi SVM dapat dipengaruhi oleh ketidakseimbangan kelas, di mana data kelas mayoritas cenderung mendominasi hasil prediksi dan mengabaikan kelas minoritas. Oleh karena itu, penelitian ini bertujuan untuk menerapkan metode *Support Vector Machine* (SVM) dalam pengklasifikasian penyakit diabetes dengan membandingkan performa model sebelum dan sesudah diterapkannya *Synthetic Minority Over-sampling Technique* (SMOTE). Data yang digunakan merupakan data PIMA Indian Diabetes dari Kaggle yang memiliki ketidakseimbangan kelas. Untuk menangani ketidakseimbangan tersebut, SMOTE digunakan dengan parameter jumlah tetangga terdekat (k) sebanyak 3 dengan label kelas positif adalah pasien terdiagnosa diabetes. Evaluasi dilakukan terhadap tiga jenis *kernel* SVM, yaitu *Linear, RBF,* dan *Polynomial*, pada data latih dan data uji. Hasil pada data latih menunjukkan bahwa model SVM dengan SMOTE memberikan peningkatan signifikan pada nilai *recall* dan *F1-score* dibandingkan model tanpa SMOTE. *Recall* mengukur kemampuan model dalam mendeteksi seluruh kasus positif (penderita diabetes), sementara *F1-score* merupakan rata-rata harmonik dari *precision* dan *recall*, yang mencerminkan keseimbangan antara kemampuan model mendeteksi dan ketepatan prediksi kasus positif. pada data latih, *kernel RBF* yang mencapai akurasi 0,9828, *recall* 0,9863, dan *F1-score* 0,9822. Sebaliknya, model tanpa SMOTE dengan *kernel RBF* menunjukkan performa yang sangat tinggi pada data latih namun rendah pada data uji, dengan *F1-score* hanya sebesar 0,1600, yang mengindikasikan kemungkinan terjadinya *overfitting*. Pada data uji, penerapan SMOTE juga menunjukkan perbaikan *recall* yang signifikan terutama pada *kernel Linear* dan *Polynomial*. Model SVM dengan SMOTE dan *kernel Linear* mencapai *recall* sebesar 0,7368 dan *F1-score* sebesar 0,6829, meningkat dibandingkan model tanpa SMOTE dengan *recall* 0,4737 dan *F1-score* 0,5806. Hal ini mengindikasikan bahwa pola klasifikasi pada data memiliki kecenderungan *linier*, sehingga transformasi *non-linear* yang ditawarkan oleh *kernel RBF* maupun *polynomial* tidak memberikan keuntungan signifikan.

Kata kunci: *Imbalance class, Kernel, Recall, F1-Score*

Application of Support Vector Machine with Synthetic Oversampling Minority Technique (SMOTE) in the Classification of Diabetes Patients

Abstract

The Support Vector Machine (SVM) algorithm is one of the most widely used algorithms in data mining for classification tasks due to its ability to handle high-dimensional data. However, the classification performance of SVM can be affected by class imbalance, where the majority class tends to dominate prediction results and overlook the minority class. Therefore, this study aims to implement the Support Vector Machine (SVM) method in classifying diabetes disease by comparing model performance before and after the application of the Synthetic Minority Over-sampling Technique (SMOTE). The dataset used is the PIMA Indian Diabetes dataset from Kaggle, which exhibits class imbalance. To address this issue, SMOTE was applied with the number of nearest neighbors (k) set to 3, where the positive class label represents patients diagnosed with diabetes. The evaluation was carried out on three types of SVM kernels: Linear, RBF, and Polynomial, using both training and testing datasets. Results on the training data showed that SVM models with SMOTE significantly improved recall and F1-score compared to models without SMOTE. Recall measures the model's ability to detect all positive cases (diabetic patients), while the F1-score is the harmonic mean of precision and recall, reflecting the balance between the model's detection ability and the accuracy of its positive predictions. On the training data, the RBF kernel achieved an accuracy of 0.9828, a recall of 0.9863, and an F1-score of 0.9822. In contrast, the model without SMOTE using the RBF kernel showed very high performance on training data but poor performance on test data, with an F1-score of only 0.1600, indicating potential overfitting. On the test data, the application of SMOTE also led to a significant improvement in recall, particularly for the Linear and Polynomial kernels. The SVM model with SMOTE and the Linear kernel achieved a recall of 0.7368 and an F1-score of 0.6829, an improvement over the model without SMOTE which achieved a recall of 0.4737 and an F1-score of 0.5806. This indicates that the classification pattern in the data tends to be linear, and the non-linear transformations offered by RBF and Polynomial kernels do not provide significant advantages.

Keywords: *Imbalance class, Kernel, Recall, F1-Score*

PRAKATA

Puji syukur kehadirat Allah SWT, Tuhan yang Maha Esa atas rahmat dan karunia-Nya penulis dapat menyelesaikan penulisan Tugas Akhir yang berjudul "Penerapan *Support Vector Machine* dengan *Synthetic Minority Oversampling* (SMOTE) Dalam Pengklasifikasian Penyakit Diabetes". Adapun penelitian ini dibuat untuk memenuhi syarat penulis untuk memperoleh gelar Sarjana Statistika pada Program Studi Statistika di Fakultas Matematika dan Ilmu Pengetahuan Alam di Universitas Tanjungpura. Penulisan ini dapat dilakukan dan diselesaikan tidak terlepas dari bimbingan, dukungan, dan saran dari berbagai pihak. Oleh karena itu, penulis berterima kasih kepada:

1. Mamak, Ibu, Abah, dan Bapak yang telah memberikan semangat, motivasi, doa dan kepercayaan pada penulis dalam menempuh pendidikan. Terima kasih penulis ucapkan juga atas dukungan moral dan materi yang dilimpahkan sehingga penulis dapat menempuh kuliah S1 hingga selesai.
2. Abang, Icu, Ncu, Fira dan seluruh keluarga besar yang telah memberikan doa, semangat dan dukungan kepada penulis selama masa perkuliahan dan penulisan skripsi ini.
3. Seluruh nama yang penulis sitasi, tanpa kalian penulis dangkal dan tersesat dalam proses penulisan skripsi ini.
4. Ibu Nurfitri Imro'ah, M.Si. selaku Dosen Pembimbing Pertama yang telah meluangkan waktu untuk membimbing, memberikan masukan dan saran kepada penulis selama proses penulisan skripsi ini.
5. Ibu Shantika Martha, M.Si. selaku Dosen Pembimbing Akademik dan Pembimbing Kedua Tugas Akhir yang telah meluangkan waktu untuk membimbing serta memberikan saran dan motivasi kepada penulis sejak awal perkuliahan hingga proses penulisan skripsi ini diselesaikan.
6. Ibu Dr. Evy Sulistianingsih, M.Sc. selaku Dosen Penguji Pertama yang telah meluangkan waktu untuk membimbing, memberikan masukan dan saran kepada penulis selama proses penulisan skripsi ini.

7. Bapak Hendra Perdana, M.Sc. selaku Dosen Penguji Kedua yang telah meluangkan waktu untuk membimbing, memberikan masukan dan saran kepada penulis selama proses penulisan skripsi ini.
8. Ummi, Wina dan Ledy yang telah membersamai, memberi dukungan dan pembelajaran hidup kepada penulis selama menjadi mahasiswa dan selalu menjaga penulis hingga mampu menyelesaikan Tugas Akhir ini.
9. Sela dan Yulia yang telah bersama penulis untuk merajut mimpi menjadi mahasiswa sejak di Sekolah Menengah Pertama, memberi dukungan dan selalu bersedia mendengarkan keluh kesah penulis selama bersama di perantauan.
10. Dayo, Ayyash, dan Seluruh Teman-teman "Expos" yang bersedia membantu, mengajarkan, berjuang bersama dan memberikan warna dalam kehidupan perkuliahan penulis selama menempuh pendidikan di Program Studi Statistika.
11. *Awardee* Untan dan Fasilitator dari BSI Scholarship yang telah memberikan kesempatan melalui materi dan motivasi kepada penulis untuk melanjutkan perkuliahan hingga proses penulisan Tugas Akhir ini diselesaikan.
12. Berbagai pihak yang telah membantu penulis dalam penyusunan skripsi ini dan belum bisa disebutkan satu per satu.

Penulis mengharapkan kritik dan saran yang bersifat membangun untuk penyempurnaan penulisan skripsi ini sehingga dapat bermanfaat bagi penulis dan para pembaca. Akhir kata penulis ucapkan terima kasih kepada semua pihak yang berperan serta selama proses penyusunan skripsi ini.

Pontianak, 6 Mei 2025

Sy. Farini Nurhaliza

DAFTAR ISI

Abstrak ..... **iii**
DAFTAR ISI ..... **vii**
DAFTAR GAMBAR ..... **ix**
DAFTAR TABEL ..... **x**
DAFTAR LAMPIRAN ..... **xi**
DAFTAR SIMBOL ..... **xii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Masalah 4
1.5 Tinjauan Pustaka 4
1.6 Metodologi Penelitian 6
BAB II LANDASAN TEORI **10**
2.1 *Diabetes Mellitus* 10
2.2 *Data Mining* 11
2.3 Machine learning 12
2.4 Ketidakseimbangan kelas data 12
2.5 *Confusion Matrix* 13
BAB III PENERAPAN SMOTE DAN KLASIFIKASI *SUPPORT VECTOR MACHINE* **16**
3.1 SMOTE 16
3.2 *Support Vector Machine* 17
BAB IV HASIL DAN PEMBAHASAN **22**
4.1 *Preprosessing* Data 22
4.2 Pembagian Data Latih dan Data Uji 26
4.3 *Support Vector machine* Tanpa *Synthetic Minority Oversampling* 27
4.3.1 Kernel Linear 27
4.3.2 Kernel RBF 29
4.3.3 Kernel Polynomial 31

4.4 *Support Vector Machine* Dengan *Synthetic Minority Oversampling*..... 33
4.4.1 *Kernel Linear*..... 34
4.4.2 *Kernel RBF* 36
4.4.3 *Kernel Polynomial* 37
4.5 Perbandingan Kinerja Model Support Vector Machine..... 39
BAB V PENUTUP 42
5.1 Kesimpulan..... 42
5.2 Saran 43
DAFTAR PUSTAKA..... 45
LAMPIRAN..... 48

DAFTAR GAMBAR

Gambar 1. 1 *Flowchart* penelitian 9
Gambar 3. 1 (a) SVM dengan *hyperplane* keputusan *linear*, dan
(b) *hyperplane* keputusan *non-linear* dengan fungsi *kernel*. 17
Gambar 3. 2 Visualisasi *kernel* dalam SVM 18
Gambar 3. 3 (a) Pemilihan parameter c yang terlalu kecil, dan
(b) Pemilihan parameter c yang terlalu besar 20
Gambar 3. 4 (a) Pemilihan parameter *gamma* yang kecil, dan
(b) Pemilihan parameter *gamma* yang baik 21
Gambar 4. 1 Deteksi *outlier* 24
Gambar 4. 2 Diagram diagnosa pasien diabetes 26
Gambar 4. 3 Perbandingan *outcome* sebelum dan sesudah SMOTE 34

DAFTAR TABEL

Tabel 1. 1 *Confusion matrix* 13
Tabel 3. 1 Fungsi *kernel support vector machine* 19
Tabel 4. 1 Variabel penelitian 22
Tabel 4. 2 Statistika deskriptif atribut independen 23
Tabel 4. 3 Perbandingan data latih dan data uji 26
Tabel 4. 4 Nilai akurasi parameter model SVM kernel linear. 27
Tabel 4. 5 Parameter SVM kernel linear tanpa SMOTE 28
Tabel 4. 6 Confusion matrix SVM kernel linear tanpa SMOTE 28
Tabel 4. 7 Nilai akurasi parameter terbaik untuk data latih model SVM *kernel RBF.* 29
Tabel 4. 8 Parameter kernel RBF tanpa SMOTE 30
Tabel 4. 9 Confusion matrix kernel RBF tanpa SMOTE 30
Tabel 4. 10 Nilai akurasi parameter terbaik dengan trial error untuk data latih Kernel Polynomial tanpa SMOTE. 31
Tabel 4. 11 Parameter kernel polynomial tanpa SMOTE 32
Tabel 4. 12 Confusion matrix kernel polynomial tanpa SMOTE 33
Tabel 4. 13 Akurasi kernel linear dengan SMOTE 34
Tabel 4. 14 Parameter SVM dengan SMOTE kernel linear 35
Tabel 4. 15 Confusion matrix SVM kernel linear dengan SMOTE 35
Tabel 4. 16 Parameter kernel RBF dengan SMOTE 36
Tabel 4. 17 Confusion matrix kernel RBF dengan SMOTE 36
Tabel 4. 18 Parameter kernel polynomial dengan SMOTE 37
Tabel 4. 19 Confusion matrix kernel polynomial dengan SMOTE 38
Tabel 4. 20 Perbandingan penerapan model SVM dengan dan tanpa SMOTE pada data latih 39
Tabel 4. 21 Perbandingan penerapan model SVM dengan dan tanpa SMOTE pada data uji 40
Tabel 5. 1 Perbandingan Kernel-kernel 43

DAFTAR LAMPIRAN

Lampiran 1 Data Latih SMOTE 47

Lampiran 2 Syntax SVM tanpa dan dengan SMOTE 48

DAFTAR SIMBOL

d : Jarak antara objek untuk mendapatkan nilai terdekat titik sampel

a_{ij} : Titik data sampel ke-i variabel ke-j , $i = 1, 2, \dots, n,\ j = 1, 2, \dots, m$

g_n : Titik perbandingan sampel ke-i variabel ke-j , $i = 1, 2, \dots, n,$ $j = 1, 2, \dots, m$

X_{new} : Nilai baru hasil SMOTE

K_{ij} : Titik sampel ke i variabel j, $i = 1, 2, \dots, n,\ j = 1, 2, \dots, m$

n : Jumlah total sampel atau observasi dalam dataset

m : Jumlah total variabel dalam setiap sampel

$f(x)$: Fungsi prediksi

$\boldsymbol{q}$: Vektor kemiringan garis model

$\boldsymbol{x_i}$: Data yang mendekati *hyperplane*

$\boldsymbol{x_j}$: Vector fitur dari data baru yang ingin diklasifikasi

b : Bias terhadap sumbu y

P : Derajat *polynomial*

γ : Derajat RBF

BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes Mellitus merupakan suatu kondisi medis kronis di mana tubuh tidak mampu mengatur kadar gula darah (glukosa) secara efektif. Penyakit diabetes adalah salah satu masalah kesehatan global yang terus meningkat dalam beberapa dekade terakhir. Diabetes menjadi salah satu penyebab utama morbiditas dan mortalitas di seluruh dunia, menurut data Riset Kesehatan Dasar (2018), di Indonesia sendiri penyakit tidak menular menjadi penyebab kematian tertinggi. Terdapat tiga jenis penyakit penyebab kematian tertinggi di Indonesia yakni stroke, serangan jantung dan *diabetes mellitus*. Berdasarkan data dari *International Diabetes Federation* (IDF), prevelensi diabetes pada tahun 2021 sebesar 573 juta orang dewasa dengan rentang umur 20-79 tahun menderita diabetes dan 19,5 juta di antaranya di Indonesia dan diperkirakan pada tahun 2030 penderita diabetes meningkat menjadi 643 juta didunia dan 28,5 juta di Indonesia. Gaya hidup tidak sehat, pola makan tinggi gula, serta kurangnya aktivitas fisik menjadi pemicu utama peningkatan jumlah penderita diabetes. Oleh karena itu, pendeteksian dini dan pengklasifikasian pasien diabetes secara akurat menjadi hal yang penting untuk mencegah komplikasi yang lebih lanjut dan mengelola risiko dengan lebih efektif. Dalam hal ini, *data mining* menjadi salah satu pendekatan yang relevan untuk mendukung pendeteksian dini diabetes.

Data mining melakukan ekstraksi informasi yang dibutuhkan dari data yang berjumlah besar dengan klasifikasi(Anugrahnu et al., 2023). Klasifikasi adalah salah satu metode analisis dalam *data mining* yang digunakan untuk memprediksi label kelas pada suatu dataset. Proses ini melibatkan dua tahap utama, yaitu tahap pembelajaran dan tahap pengklasifikasian. Menurut Sartika & Sensuse (2017), tahap pembelajaran merupakan tahapan dalam membentuk model klasifikasi, sedangkan tahap pengklasifikasian merupakan tahap penggunaan model klasifikasi untuk memprediksi label kelas dari suatu data. Dalam konteks diabetes, klasifikasi melalui data mining sangat membantu dalam memprediksi apakah seseorang

berisiko atau tidak terkena diabetes. Kondisi yang kompleks dan melibatkan faktor seperti kadar glukosa darah, insulin dan faktor genetik yang semuanya memerlukan analisis yang cermat untuk diagnosis yang tepat. Dalam konteks pengklasifikasian diabetes, SVM atau *Support Vector Machine* sangat relevan karena kemampuannya dalam menangani data yang kompleks dan *multidimensional*. SVM bekerja dengan membangun *hyperplane* yang memisahkan data dari dua kelas (Terdiagnosa Diabetes dan Tidak Terdiagnosa Diabetes) secara optimal berdasarkan fitur-fitur medis yang relevan. SVM adalah model tunggal algoritma *non-ensemble* yang mencari *hyperplane* terbaik untuk memisahkan data menjadi kelas yang berbeda. Dalam penelitian Mohammed, *et al* (2016) menyatakan SVM dirancang oleh algoritma pembelajaran mesin dengan pendekatan pembelajaran *supervised* yang dijalankan untuk menemukan fungsi pemisah terbaik dalam memisahkankan kelas. SVM bertujuan meminimalkan batas atas dari kesalahan dalam pengukuran. Metode ini juga menggunakan konsep teori pembelajaran komputasi yang mengembangkan algoritma prediksi atau mengklasifikasikan data baru berdasarkan contoh-contoh atau data pelatihan yang diberikan.

Berdasarkan data medis, sejumlah besar pasien dapat dikelompokkan sebagai pasien dengan dan tanpa diabetes. Namun seringkali terjadi ketidakseimbangan dalam jumlah pasien di setiap kelompok, dimana lebih banyak pasien yang masuk ke kelompok tanpa diabetes. Hal ini membuat sulit untuk mengidentifikasi pola dan faktor yang mempengaruhi terjadinya diabetes pada kelompok minoritas tersebut. Ketidakseimbangan kelas dapat mengakibatkan data dari kelompok risiko tinggi menjadi terabaikan oleh model analisis *Support Vector Machine*. Untuk mengatasi masalah ini, diterapkan teknik *Synthetic Minority Oversampling Technique* (SMOTE), sebuah teknik dalam *machine learning* yang dirancang untuk menangani ketidakseimbangan kelas. SMOTE bekerja dengan membuat sampel sintetik dari kelompok minoritas, dalam hal ini pasien dengan risiko tinggi terkena diabetes, dengan cara menginterpolasi data yang ada. Sampel sintetis ini memiliki karakteristik yang mirip dengan data asli sehingga dapat memperkaya data kelas minoritas tanpa menambah bias. SMOTE secara efektif

meningkatkan jumlah sampel pada kelas minoritas dengan cara mensitesis data baru berdasarkan sampel yang sudah ada.

Dengan menggunakan SMOTE, jumlah sampel dari pasien berisiko tinggi yang awalnya sedikit akan bertambah, sehingga memungkinkan model SVM untuk mempelajari pola-pola risiko diabetes dengan lebih baik. Hal ini akan meningkatkan akurasi, *recall* dan *specificity*, terutama untuk pasien dengan risiko tinggi yang sebelumnya mungkin sulit dikenali. Kombinasi antara SMOTE dan SVM dapat membantu membangun model klasifikasi yang lebih kuat dan akurat, yang pada gilirannya dapat memberikan wawasan yang lebih jelas mengenai faktor-faktor penting yang mempengaruhi risiko diabetes. Tujuannya penelitian ini dapat memberikan kontribusi dalam mengupayakan pencegahan dan mengendalikan *diabetes mellitus*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan disusunlah rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana penerapan metode *Support Vector Machine* dalam mengklasifikasikan pasien diabetes?
2. Bagaimana penerapan *Synthetic Minority Oversampling Technique* (SMOTE) dapat membantu menangani ketidakseimbangan kelas pada data pasien diabetes?
3. Apakah penerapan SMOTE dapat meningkatkan kinerja klasifikasi SVM dalam mengatasi ketidakseimbangan kelas pada data pasien diabetes?

1.3 Tujuan Penelitian

Tujuan penelitian penerapan *Support Vector Machine* dengan SMOTE dalam pengklasifikasi pasien dengan dan tanpa diabetes dapat dirumuskan sebagai berikut.

1. Menerapkan Metode *Support Vector Machine* dalam mengklasifikasikan pasien diabetes.
2. Menerapkan Metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk menangani ketidakseimbangan kelas pada data pasien diabetes.

3. Menerapkan SMOTE untuk meningkatkan kinerja klasifikasi SVM dalam mengatasi ketidakseimbangan kelas pada data pasien diabetes.

1.4 Batasan Masalah

Batasan masalah dalam penelitian penerapan *Support Vector Machine* (SVM) dengan SMOTE untuk mengklasifikasi pasien dengan dan tanpa diabetes ditetapkan sebagai berikut:

1. Data yang digunakan dalam penelitian ini berasal dari kumpulan data *Pima Indians Diabetes* yang tersedia di Kaggle tahun 2016.
2. Penelitian ini menggunakan variabel dependen berupa *outcome*, yaitu status pasien yang dikategorikan sebagai tanpa diabetes dan dengan diabetes. Variabel independen yang digunakan meliputi Kehamilan (*Pregnancies*), Kadar Gula Darah (*Glucose*), Tekanan Darah (*Blood pressure*), Ketebalan Lipatan Kulit (*Skin thickness*), Kadar Insulin (Insulin), Indeks Massa Tubuh (BMI), Riwayat Keluarga Diabetes (*Diabetes pedigree function*), dan Usia (*Age*). Data yang digunakan akan dibagi menjadi data latih dan data uji dengan perbandingan 90:10.
3. *Kernel* SVM yang digunakan dalam penelitian ini hanya *kernel linear, kernel polynomial,* dan *kernel RBF*.

1.5 Tinjauan Pustaka

Chawla et al. (2002) memperkenalkan *Synthetic Minority Over-sampling Technique* (SMOTE) sebagai metode yang mampu mengatasi masalah ketidakseimbangan kelas dalam dataset, khususnya pada pengklasifikasian *machine learning.* SMOTE bekerja dengan cara mensintesis data baru untuk kelas minoritas, sehingga menciptakan distribusi kelas yang lebih seimbang dan meningkatkan kinerja model klasifikasi. Mekanisme kerja SMOTE melibatkan pemilihan sampel terdekat dari kelas minoritas, perhitungan vektor fitur baru dan penambahan data sintetik kedalam dataset. Parameter *k* dalam SMOTE memainkan peran penting untuk menentukan jumlah tetangga terdekat yang digunakan untuk menghasilkan data sintetik. Nilai *k* yang optimal akan bergantung pada karakteristik dataset yang berbeda-beda.

Pada penelitian Patil dan Nemade tahun 2017 mengklasifikasikan jenis musik menggunakan model MFCC yang merupakan sekumpulan karakteristik spektrum daya jangka pendek dari file audio, penelitian ini mempertimbangkan 13 koefisien sebagai bagian dan vektor fitur akhir. Setelah vektor fitur diperoleh, dilatih klasifikator berbeda yakni *K-Nearest Neighbors, Linear kernel* SVM dan *Polynomial Kernel* SVM pada kumpulan data latih vektor fitur tersebut. Penelitian ini mengukur kinerja sistem menggunakan metrik akurasi, recall dan presisi. Semakin tinggi nilai *recall* dan *presisi,* akan memberi efisiensi yang lebih baik dalam klasifikasi. Hasil penelitian menunjukan bahwa SVM dengan *Kernel polynomial* menjadi pengklasifikasi terbaik dengan nilai akurasi sebesar 78%, nilai presisi sebesar 79% dan nilai *recall* sebesar 78%.

Hidayat et al (2021) dalam penelitiannya mengenai klasifikasi *super host Airbnb* menerapkan SVM dengan algoritma ADASYN dan SMOTE memberikan keseimbangan hasil yang lebih baik. ADASYN meningkatkan kesetaraan hasil pengujian antara label *True* dan *False*, tetapi Presisi dan *F1-score* untuk *label false* mengalami penurunan. Sedangkan, penerapan SMOTE meningkatkan Presisi, *Recall, F1-SCORE* dan Akurasi yang lebih baik dibandingkan ADASYN. Penerapan SMOTE pada algoritma SVM meningkatkan akurasi klasifikasi sebesar 1% dibandingkan dengan ADASYN SVM dan SVM tanpa *oversampling.*

Selanjutnya penelitian klasifikasi penyakit diabetes menggunakan metode *Support Vector Machine* yang dilakukan oleh Mucholladin et al. (2021). Tujuan penelitian ini adalah membuat model *machine learning* yang dapat mendeteksi dini penyakit diabetes. *Support Vector Machine* (SVM) adalah salah satu metode *machine learning* yang dikenal cukup efektif untuk kasus klasifikasi. Dataset dibersihkan dan dinormalisasi terlebih dahulu sehingga siap untuk dimasukkan ke dalam model SVM. Model SVM diproses dan diuji sehingga mendapatkan model terbaik untuk melakukan diagnosis. Keluaran dari model SVM akan mendiagnosis pasien yang menderita diabetes ataupun yang tidak menderita diabetes. Pada penelitian ini dilakukan penanganan anomali untuk melihat fitur yang memiliki *outlier*. Dalam menangani *Outlier,* Hasil pengujian menunjukkan bahwa model benchmark memiliki nilai 0,87 mean *accuracy*, 0,82 *mean precision*, 0,78 *mean*

sensitivity, dan 0,92 *mean specificity*. Model scratch memiliki nilai 0,78 *mean accuracy*, 0,69 *mean precision*, 0,59 *mean sensitivity*, dan 0,87 *mean specificity*. Hasil eksperimen menunjukkan bahwa metode *Support Vector Machin*e memiliki potensi untuk digunakan sebagai alat deteksi dini penyakit diabetes.

Pada penelitian Lestari dan Aryanto, (2023) yang bertujuan untuk meningkatkan model klasifikasi kualitas udara menggunakan metode *Random Forest* dan *Support Vector Machine* dengan penerapan teknik *oversampling* SMOTE. Data yang digunakan adalah Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta selama tahun 2022. Pada pemodelan dengan *Support Vector Machine*, akurasi mencapai 91%, namun dengan SMOTE, akurasi meningkat menjadi 95%.

1.6 Metodologi Penelitian

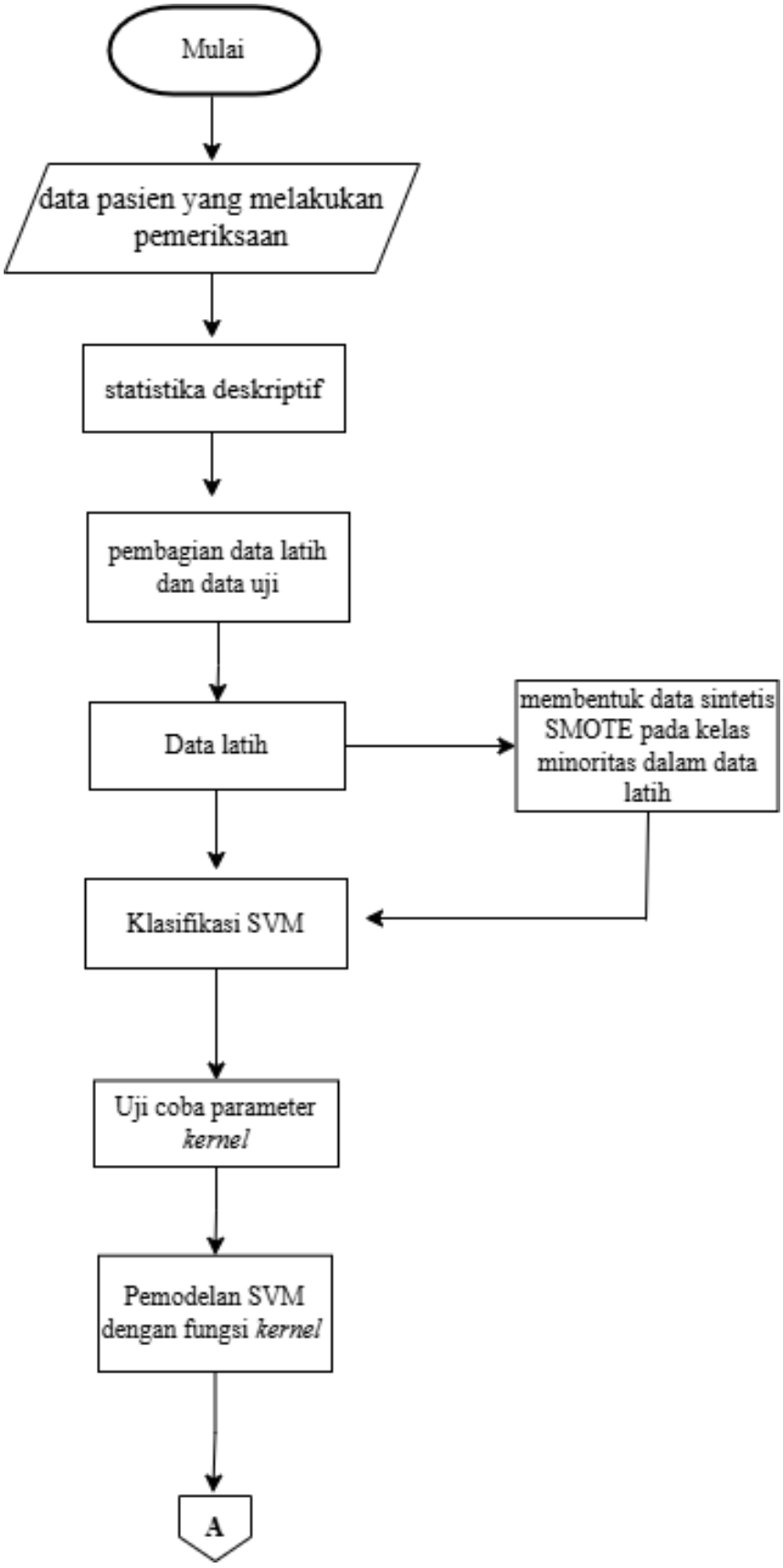
Metode penelitian yang digunakan dalam penelitian ini menggunakan pendekatan eksperimental untuk menganalisis data pasien yang didiagnosis dengan diabetes. Variabel-variabel yang dianalisis dalam penelitian ini mencakup berbagai aspek kesehatan, antara lain kadar glukosa darah sebagai indikator utama, indeks massa tubuh (BMI) untuk menggambarkan proporsi berat badan terhadap tinggi badan, tekanan darah sebagai salah satu faktor risiko kardiovaskular, usia pasien yang dapat memengaruhi risiko diabetes, serta riwayat keluarga yang menggambarkan predisposisi genetik terhadap penyakit ini. Selain itu, data mencakup riwayat kehamilan untuk wanita, kadar insulin sebagai indikator regulasi gula darah, dan ketebalan kulit yang dapat mencerminkan *sensitivitas* insulin. Data yang diperoleh selanjutnya diolah dan dipersiapkan untuk dianalisis lebih lanjut. Proses pengolahan data ini mencakup pembagian dataset menjadi dua bagian utama, yaitu data latih dan data uji, dengan tujuan untuk memastikan model yang dihasilkan memiliki kemampuan generalisasi yang baik. Data latih terdiri dari 90% dari keseluruhan data, sementara data uji mencakup 10% sisanya.

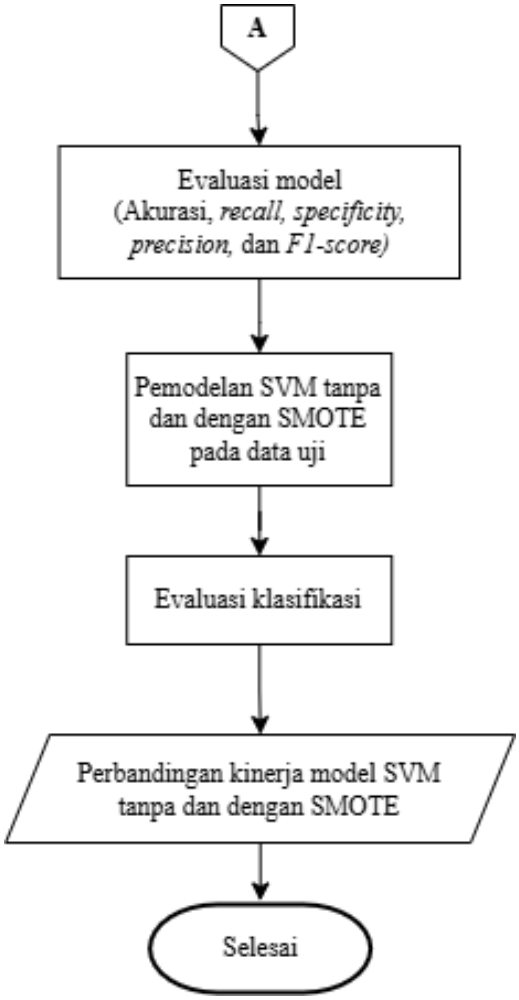
Algoritma Support Vector Machine (SVM) diterapkan dengan menguji tiga jenis kernel yakni *linear, RBF,* dan *polynomial*. Pada *kernel linear,* parameter C (Regularization Parameter) diuji untuk mengontrol keseimbangan antara margin dan salah klasifikasi. Pada *kernel RBF*, parameter C dan gamma (γ) diuji, di mana gamma mengontrol pengaruh jarak data terhadap keputusan klasifikasi. Untuk

kernel polynomial, parameter yang diuji meliputi C, dan *degree* (derajat polinomial). Proses pengujian parameter dilakukan menggunakan *trial error* parameter terbaik dipilih berdasarkan nilai *error* terkecil. Kombinasi parameter dioptimalkan menggunakan *grid search* atau *random search*, dengan evaluasi berbasis *confusion matrix, recall, precision* dan *specificity* untuk melihat apakah SVM tanpa SMOTE mampu menangani ketidakseimbangan kelas. Apabila *confusion matrix, recall, precision* dan *specificity* menunjukkan ketidakseimbangan kelas maka akan diterapkan model SVM dengan SMOTE untuk setiap *kernel.*

Untuk menangani masalah *imbalanced class data,* SMOTE diterapkan pada data latih, sehingga jumlah pasien dalam kelas minoritas (misalnya pasien dengan diabetes) mendekati jumlah dalam kelas mayoritas (pasien tanpa diabetes). Setelah pra-pemroresan data, dilakukan penyeimbangan data guna meningkatkan akurasi dan mengatasi ketidakseimbangan antara jumlah pasien yang terdiagnosis diabetes dengan yang tidak. Kelas minor yang telah ditambahkan menghasilkan data yang disebut dengan data buatan. Pembangkitan data untuk data numerik dihitung perbedaan antara vektor utama dengan K tetangga terdekat kemudian mengalikan perbedaan tersebut dengan angka yang diacak pada rentang 0 dan 1 sehingga memperoleh vektor utama baru dengan menambahkan perbedaan kedalam nilai utama pada vektor utama asal.

Kemudian algoritma SVM diterapkan kembali menggunakan parameter yang sama dengan parameter *trial error* sebelum SMOTE. Kemudian mengevaluasi kembali kinerja model SVM menggunakan *confusion matrix*, *recall, precision* dan *specificity*. Langkah terakhir, model SVM dengan fungsi *kernel* yang optimal diterapkan pada data uji untuk mengukur akurasi klasifikasi diabetes. Proses keseluruhan dijelaskan dalam bentuk diagram alir yang ditampilkan pada Gambar 1.1.


Mulai
data pasien yang melakukan pemeriksaan
statistika deskriptif
pembagian data latih dan data uji
Data latih
membentuk data sintetis SMOTE pada kelas minoritas dalam data latih
Klasifikasi SVM
Uji coba parameter *kernel*
Pemodelan SVM dengan fungsi *kernel*
A



Gambar 1. 1 *Flowchart* penelitian