



**INTEGRASI SMOTE-NC DAN SUPPORT VECTOR MACHINE
UNTUK PENANGANAN DATA TIDAK SEIMBANG PADA
KLASIFIKASI PENDERITA DIABETES: ANALISIS SKI 2023**

SKRIPSI

**untuk memenuhi persyaratan
dalam menyelesaikan program sarjana Strata-1 Statistika**

**Oleh
AGNES ANGELINE HUTAGALUNG
NIM. 2211017220005**

**PROGRAM STUDI S-1 STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JUNI 2026**

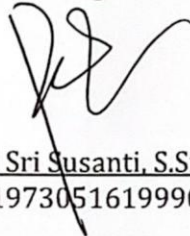
SKRIPSI

INTEGRASI SMOTE-NC DAN SUPPORT VECTOR MACHINE UNTUK PENANGANAN DATA TIDAK SEIMBANG PADA KLASIFIKASI PENDERITA DIABETES: ANALISIS SKI 2023

Oleh
Agnes Angeline Hutagalung
NIM. 2211017220005

Telah dipertahankan pada hari Rabu tanggal 10-06-2026 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

Pembimbing I



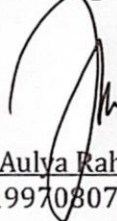
Dewi Sri Susanti, S.Si., M.Si.
NIP. 197305161999032002

Penguji I



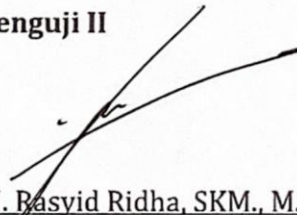
Selvi Annisa, S.Si., M.Si.
NIP. 199212262022032016

Pembimbing II



Rifqi Aulya Rahman, S.Mat., M.Si.
NIP. 199708072024061003

Penguji II



M. Basyid Ridha, SKM., M.Si.
NIP. 198404072006041003

Banjarbaru, Juni 2026
Mengetahui,
Koordinator Program Studi Statistika
FMIPA ULM



Dewi Anggraini, S.Si., M.App.Sci., Ph.D.
NIP. 198303282005012001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 3 Juni 2026



Agnes Angeline Hutagalung

NIM. 2211017220005

ABSTRAK

Integrasi SMOTE-NC dan *Support Vector Machine* untuk Penanganan Data Tidak Seimbang pada Klasifikasi Penderita Diabetes: Analisis SKI 2023

(Oleh: Agnes Angeline Hutagalung; Pembimbing: Dewi Sri Susanti dan Rifqi Aulya Rahman, 2026; 48 halaman)

Ketidakseimbangan data merupakan salah satu permasalahan dalam klasifikasi yang menyebabkan model cenderung mengabaikan kelas minoritas. Kondisi ini umum terjadi pada data kesehatan yang memuat distribusi kelas yang tidak proporsional, seperti pada data penderita diabetes. Jika dilakukan klasifikasi pada data tersebut berpotensi menurunkan kemampuan model dalam menghasilkan prediksi yang optimal. Oleh karena itu, diperlukan metode penanganan ketidakseimbangan data yang sesuai, khususnya untuk karakteristik data dengan kombinasi variabel kategorik dan numerik. Penelitian ini bertujuan mengaplikasikan metode SMOTE-NC untuk mengatasi ketidakseimbangan data campuran, serta menganalisis kinerjanya pada klasifikasi dengan algoritma SVM menggunakan tiga jenis kernel (*linear*, *polynomial*, RBF) dengan variasi parameter *cost* (C) dan gamma (γ). Metode penelitian menggunakan data sekunder Survei Kesehatan Indonesia (SKI) 2023 dengan prosedur meliputi *preprocessing* data, pembagian data, penyeimbangan data latih menggunakan SMOTE-NC, serta pengujian model SVM melalui *manual tuning* parameter. Hasil penelitian menunjukkan sebelum penyeimbangan, model SVM sama sekali tidak mampu mendeteksi kelas minoritas. Setelah penerapan SMOTE-NC, kinerja model meningkat signifikan dalam mendeteksi kelas minoritas, dengan kernel *linear* ($C = 1$) dan kernel RBF ($C = 0.1$; $\gamma = 0.1$) sebagai model terbaik. Kesimpulannya, integrasi SMOTE-NC meningkatkan kinerja model SVM dalam mendeteksi kelas minoritas, meskipun masih terdapat kelemahan pada nilai *precision* akibat tingginya *false positive*.

Kata Kunci: Diabetes Melitus, Ketidakseimbangan Data, SMOTE-NC, *Support Vector Machine*

ABSTRACT

Integration of SMOTE-NC and Support Vector Machines for Handling Imbalanced Data in Diabetes Patient Classification: SKI 2023 Analysis

(By: Agnes Angeline Hutagalung; Supervisors: Dewi Sri Susanti and Rifqi Aulya Rahman, 2026; 48 pages)

Data imbalanced is a common issue in classification that causes models to neglect the minority class. This condition is common in health dataset containing disproportionate class distributions, such as diabetes patient data. Classifying such data has the potential to reduce the model's ability to generate optimal predictions. Therefore, appropriate imbalance-handling methods are needed, especially for datasets containing a mix of categorical and numerical variables. This study aims to apply the SMOTE-NC method to address mixed-data imbalance and to analyze its performance on SVM algorithms using three kernel types (linear, polynomial, RBF) with variations in cost (C) and gamma (γ) parameters. The research used secondary data from the 2023 Indonesian Health Survey (Survei Kesehatan Indonesia, SKI) with procedures including data preprocessing, data splitting, training-set balancing using SMOTE-NC, and SVM model evaluation through manual parameter tuning. Results show that before balancing, the SVM models were unable to detect the minority class. After applying SMOTE-NC, model performance improved substantially in detecting the minority class, with the linear kernel ($C = 1$) and the RBF kernel ($C = 0.1$; $\gamma = 0.1$) identified as the best models. In conclusion, the integration of SMOTE-NC improves the performance of the SVM model in detecting the minority class, although precision remained limited due to a high number of false positives.

Keywords: Diabetes Mellitus, Imbalanced Data, SMOTE-NC, Support Vector Machine

PRAKATA

Segala puji syukur penulis panjatkan kepada Tuhan Yesus Kristus, karena hanya oleh kasih karunia, berkat, dan penyertaan-Nya yang melimpah, penulis dapat menyelesaikan Tugas Akhir yang berjudul “Integrasi SMOTE-NC dan *Support Vector Machine* untuk Penanganan Data Tidak Seimbang pada Klasifikasi Penderita Diabetes: Analisis SKI 2023”. Tugas Akhir ini disusun untuk memenuhi salah satu persyaratan dalam menyelesaikan program sarjana Strata-1 di Program Studi Statistika FMIPA ULM. Penulis ingin mengucapkan terima kasih yang

1. Kedua orang tua dan keluarga tercinta yang selalu memberikan doa, dukungan, dan motivasi kepada penulis agar senantiasa mengandalkan Tuhan dalam segala hal, serta percaya bahwa penulis mampu melewati proses ini;
2. Ibu Dewi Sri Susanti, S.Si., M.Si. dan Bapak Rifqi Aulya Rahman, S.Mat., M.Si. selaku dosen pembimbing yang telah memberikan bimbingan dan arahan dalam proses penyusunan Tugas Akhir;
3. Ibu Selvi Annisa, S.Si., dan Bapak M. Rasyid Ridha, SKM., M.Si selaku dosen penguji yang telah meluangkan waktu dan memberikan masukan, saran, serta bantuan dalam proses perbaikan penulisan Tugas Akhir ini;
4. Ibu Dewi Anggraini, S.Si., M.App.Sci., Ph.D selaku Koordinator Program Studi Statistika, serta Bapak/Ibu Dosen dan Staff Program Studi Statistika yang telah memberikan ilmu, arahan, serta motivasi yang sangat berharga selama penulis menempuh masa perkuliahan;
5. Sahabat dan teman-teman terdekat yang senantiasa menemani, mendengarkan, dan memberikan dukungan serta nasihat selama penulis menyelesaikan Tugas Akhir ini;
6. Teman-teman *Funtastic* atas segala kenangan berharga dan perjuangan bersama yang telah terlewati sejak awal masa perkuliahan;
7. Pihak-pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan dukungan, serta motivasi yang berarti untuk penulis selama penulisan Tugas Akhir ini.

Penulis tentu menyadari masih terdapat kekurangan dalam penulisan Tugas Akhir ini, sehingga kritik dan saran yang bertujuan membangun sangat diharapkan untuk membantu menyempurnakan Tugas Akhir ini. Akhir kata, penulis berharap semoga hasil penelitian ini bermanfaat bagi semua pihak. Kasih karunia dan damai sejahtera Tuhan menyertai kita semua.

Banjarbaru, Juni 2026

Agnes Angeline Hutagalung

DAFTAR ISI

LEMBAR PENGESAHAN	ii
PERNYATAAN	iii
ABSTRAK.....	iv
ABSTRACT	v
DAFTAR ISI	vii
DAFTAR GAMBAR	ix
DAFTAR TABEL	x
DAFTAR LAMPIRAN	xi
DAFTAR LAMBANG DAN SINGKATAN.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian.....	3
BAB II TINJAUAN PUSTAKA	4
2.1 Kajian Penelitian Terdahulu.....	4
2.2 Kajian Teori	6
2.2.1 Statistika Deskriptif.....	6
2.2.2 Ketidakseimbangan Data.....	6
2.2.3 Penanganan Ketidakseimbangan Data.....	7
2.2.4 <i>Data Mining</i>	9
2.2.5 <i>Splitting Data</i>	10
2.2.6 <i>Support Vector Machine (SVM)</i>	11
2.2.7 <i>Hyperparameter Tuning</i>	14
2.2.8 Evaluasi Model.....	15
2.2.9 Diabetes Melitus	17
BAB III METODE PENELITIAN.....	21
3.1 Sumber Data.....	21
3.2 Variabel Penelitian.....	21
3.3 Prosedur Penelitian.....	22
BAB IV HASIL DAN PEMBAHASAN	26
4.1 <i>Preprocessing</i> Data.....	26
4.2 Statistika Deskriptif.....	27
4.3 <i>Splitting Data</i>	33
4.4 Pemodelan SVM Sebelum SMOTE-NC (<i>Baseline</i>)	34
4.5 Evaluasi Model SVM Sebelum SMOTE-NC	34
4.6 Penerapan SMOTE-NC.....	38
4.7 Pemodelan SVM Setelah SMOTE-NC.....	39
4.8 Evaluasi Model SVM Setelah SMOTE-NC.....	39
4.9 Perbandingan Kinerja Model Sebelum dan Sesudah SMOTE-NC ..	45
BAB V PENUTUP	47
5.1 Kesimpulan.....	47
5.2 Saran	48
DAFTAR PUSTAKA	49

LAMPIRAN.....	54
RIWAYAT HIDUP	96

DAFTAR GAMBAR

Gambar 2.1	Ilustrasi <i>Hyperplane</i>	11
Gambar 2.2	(a) <i>Cost</i> Tinggi (b) <i>Cost</i> Rendah.....	13
Gambar 2.3	(a) Γ Tinggi (b) Γ Rendah.....	13
Gambar 3.1	Prosedur Analisis.....	25
Gambar 4.1	Visualisasi <i>Missing Value</i> pada Data	26
Gambar 4.2	Proporsi Diagnosa Diabetes	28
Gambar 4.3	Distribusi Diagnosa Diabetes Berdasarkan Status Merokok	29
Gambar 4.4	Distribusi Diagnosa Diabetes Berdasarkan Aktivitas Fisik	30
Gambar 4.5	Distribusi Diagnosa Diabetes Berdasarkan Pendidikan Terakhir	31
Gambar 4.6	Sebaran Data Sebelum dan Sesudah SMOTE-NC (<i>Data Training</i>)	38
Gambar 4.7	Kinerja Model SVM Kernel <i>Linear</i> Berdasarkan Nilai C dan k ..	43
Gambar 4.8	Kinerja Model SVM Kernel RBF Berdasarkan Nilai C , γ , dan k ..	44
Gambar 4.9	Kinerja Model SVM Terbaik pada Kernel <i>Linear</i>	45
Gambar 4.10	Kinerja Model SVM Terbaik pada Kernel RBF	46

DAFTAR TABEL

Tabel 2.1	Kajian Penelitian Terdahulu.....	4
Tabel 2.2	<i>Confusion Matrix</i>	15
Tabel 2.3	Literatur Pemilihan Variabel.....	18
Tabel 3.1	Variabel Penelitian.....	21
Tabel 4.1	Statistika Deskriptif Konsumsi Buah Berdasarkan Diagnosa Diabetes.....	30
Tabel 4.2	Statistika Deskriptif Konsumsi Sayur Berdasarkan Diagnosa Diabetes.....	32
Tabel 4.3	Statistika Deskriptif Kesehatan Mental Berdasarkan Diagnosa Diabetes.....	32
Tabel 4.4	Statistika Deskriptif Umur Berdasarkan Diagnosa Diabetes.....	32
Tabel 4.5	Statistika Deskriptif BMI Berdasarkan Diagnosa Diabetes.....	33
Tabel 4.6	Hasil <i>Splitting Data</i>	34
Tabel 4.7	<i>Confusion Matrix</i> SVM Sebelum SMOTE-NC	35
Tabel 4.8	Hasil <i>Tuning Parameter</i> SVM Kernel <i>Linear</i>	36
Tabel 4.9	Hasil <i>Tuning Parameter</i> SVM Kernel RBF	37
Tabel 4.10	Jumlah dan Persentase Data Sebelum dan Sesudah SMOTE-NC (<i>Data Training</i>).....	38
Tabel 4.11	<i>Confusion Matrix</i> SVM Kernel <i>Linear</i> Setelah SMOTE-NC.....	40
Tabel 4.12	Hasil <i>Tuning Parameter</i> SVM Kernel <i>Linear</i> Setelah SMOTE-NC.....	41
Tabel 4.13	<i>Confusion Matrix</i> SVM Kernel RBF Setelah SMOTE-NC	41
Tabel 4.14	Hasil <i>Tuning Parameter</i> SVM Kernel RBF Setelah SMOTE-NC	43
Tabel 4.15	Perbandingan Kinerja Model SVM Terbaik untuk Tiap Kernel....	45

DAFTAR LAMPIRAN

Lampiran 1.	Surat Pengambilan dan Pengambilan Data.....	54
Lampiran 2.	<i>Dataset</i> Sebelum Dilakukan <i>Prerocessing</i>	55
Lampiran 3.	<i>Dataset</i> Setelah Dilakukan <i>Prerocessing</i>	55
Lampiran 4.	<i>Dataset</i> pada Data <i>Training</i>	55
Lampiran 5.	<i>Dataset</i> pada Data <i>Testing</i>	56
Lampiran 6.	Syntax Python untuk <i>Import Data</i> dan <i>Preprocessing</i>	56
Lampiran 7.	Syntax Python untuk Statistika Deskriptif.....	59
Lampiran 8.	Syntax Python <i>Spliting Data</i> dan Klasifikasi Model SVM Sebelum SMOTE-NC	65
Lampiran 9.	Syntax Python Klasifikasi Model SVM Dengan SMOTE-NC (k=4).....	72
Lampiran 10.	Syntax Python Klasifikasi Model SVM Dengan SMOTE-NC (k=5).....	77
Lampiran 11.	Syntax Python Klasifikasi Model SVM Dengan SMOTE-NC (k=6).....	83
Lampiran 12.	Syntax Python Visusalisasi Perbandingan Hasil $k = 4, 5, 6$	89
Lampiran 13.	Syntax Python Visusalisasi Sebaran Data Latih Sebelum dan Sesudah SMOTE-NC	92
Lampiran 14.	Syntax Python Visualisasi Perbandingan Kinerja Model SVM Terbaik untuk Setiap Kernel.....	94
Lampiran 15.	<i>Confusion</i> Matrix SVM Sebelum SMOTE-NC untuk Setiap Kombinasi.....	96
Lampiran 16.	<i>Confusion</i> Matrix SVM Setelah SMOTE-NC untuk Setiap Kombinasi.....	96

DAFTAR LAMBANG DAN SINGKATAN

C	<i>Cost</i>
γ	<i>Gamma</i>
SMOTE	<i>Synthetic Minority Oversampling Technique</i>
SMOTE-NC	<i>Synthetic Minority Oversampling Technique-Nominal Continuous</i>
SVM	<i>Support Vector Machine</i>
RBF	<i>Radial Basis Function</i>
KDD	<i>Knowledge Discovery from Data</i>
IR	<i>Imbalance Ratio</i>
SKI	<i>Survei Kesehatan Indonesia</i>