



**PENERAPAN RFECV DAN SMOTE-NC DALAM MODEL *RANDOM FOREST* UNTUK KLASIFIKASI INDIKASI DIABETES**

**SKRIPSI**

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

**Oleh  
NOVIA RAMADHANI PUTRI ZASKIA  
NIM. 2111017220003**

**PROGRAM STUDI S-1 STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
FEBRUARI 2025**



**PENERAPAN RFECV DAN SMOTE-NC DALAM MODEL *RANDOM FOREST* UNTUK KLASIFIKASI INDIKASI DIABETES**

**SKRIPSI**

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

**Oleh  
NOVIA RAMADHANI PUTRI ZASKIA  
NIM. 2111017220003**

**PROGRAM STUDI S-1 STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
FEBRUARI 2025**

## SKRIPSI

### **PENERAPAN RFECV DAN SMOTE-NC DALAM MODEL *RANDOM FOREST* UNTUK KLASIFIKASI INDIKASI DIABETES**

Oleh  
**Novia Ramadhani Putri Zaskia**  
**NIM. 2111017220003**

Telah dipertahankan pada hari Kamis, tanggal 13 Februari 2025 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

#### **Pembimbing I**



Fuad Muhajirin Farid, S.Pd., M.Si  
NIP. 198807112019031014

#### **Penguji I**



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

#### **Pembimbing II**



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

#### **Penguji II**



Oni Soesanto, S.Si., M.Si  
NIP. 197301262005011003

Banjarbaru, 25 Februari 2025

Mengetahui,

Koordinator Program Studi Statistika  
PMIPA ULM



Prof. 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.

Baniarbaru, 25 Februari 2025



Novia Ramadhani Putri Zaskia  
NIM. 2111017220003

## ABSTRAK

**Penerapan RFECV dan SMOTE-NC dalam Model *Random Forest* untuk Klasifikasi Indikasi Diabetes** (Oleh: Novia Ramadhani Putri Zaskia; Pembimbing: Fuad Muhajirin Farid dan Selvi Annisa, 2025; 83 halaman)

Teknologi untuk mengumpulkan dan menyajikan data berkembang semakin pesat sehingga penyimpanan data semakin efisien. Hal ini membuat data memiliki ukuran dimensi yang lebih besar, salah satunya yaitu data indikasi penyakit diabetes dari survei BRFSS Amerika Serikat. Data berdimensi besar seringkali memiliki sebagian variabel prediktor yang kurang berkontribusi terhadap prediksi variabel target. Selain itu, data berdimensi besar sering mengalami masalah *imbalance class* atau ketidakseimbangan kelas yang dapat mengganggu proses analisis, terutama analisis klasifikasi. Kedua masalah ini dapat diatasi dengan penerapan seleksi fitur RFECV dan teknik *resampling* SMOTE-NC. Penelitian ini bertujuan untuk mengimplementasikan RFECV dan SMOTE-NC dalam model *random forest* dan mengetahui performa kinerja model dalam mengklasifikasikan indikasi diabetes. Pembentukan model diawali dengan membangun model *random forest* menggunakan data *training*, kemudian menerapkan model tersebut ke dalam algoritma RFECV untuk mengeliminasi variabel yang kurang berpengaruh. Analisis dilanjutkan dengan *me-resampling* data *training* menggunakan SMOTE-NC, kemudian melatih kembali model *random forest* menggunakan data *training* baru. Proses dilanjutkan dengan menguji model menggunakan data *testing* dan mengevaluasinya. Dari hasil evaluasi, diperoleh performa model dengan nilai sensitivitas sebesar 0.64393, spesifisitas sebesar 0.76079, *balanced accuracy* sebesar 0.70236, dan nilai AUC sebesar 0.77489 dengan tingkat Klasifikasi Cukup.

Kata Kunci: *Imbalance Class*, RFECV, SMOTE-NC, *Random Forest*, Indikasi Penyakit Diabetes

## **ABSTRACT**

**Implementation of RFECV and SMOTE-NC in Random Forest Model for Diabetes Indication Classification** (By: Novia Ramadhani Putri Zaskia; Supervisors: Fuad Muhajirin Farid and Selvi Annisa; 83 page)

Technology for data collection and presentation is advancing rapidly, making storage increasingly efficient. This expands data dimensions, including diabetes indication data from the U.S. BRFSS survey. Large-dimensional data often includes predictor variables that do not contribute to target variable prediction. Additionally, large-dimensional data often face imbalanced class issues, which can disrupt the classification analysis process. These issues can be addressed using RFECV and SMOTE-NC. This study aims to implement RFECV and SMOTE-NC in a random forest model and evaluate its performance in classifying indications of diabetes. Model building begins with building a random forest using training data, then applying RFECV to eliminate less influential variables. The analysis continues with SMOTE-NC resampling of the training data, followed by retraining the random forest model using the new training data. The process proceeds with model testing using the testing data, followed by evaluation. From the evaluation results, the model performance was obtained with a sensitivity value of 0.64393, a specificity of 0.76079, a balanced accuracy of 0.70236, and an AUC value of 0.77489 with a Fair Classification level.

Keywords: Imbalance Class, RFECV, SMOTE-NC, Random Forest, Diabetes Indication

## PRAKATA

Alhamdulillah robbil 'alamin, puji dan syukur penulis ucapkan ke hadirat Allah SWT yang telah memberikan rahmat, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "Penerapan RFECV dan SMOTE-NC dalam Model *Random Forest* untuk Klasifikasi Indikasi Diabetes". Tugas Akhir ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan program sarjana strata-1 di Program Studi Statistika FMIPA ULM.

Dalam penyusunan Tugas Akhir ini, penulis mendapatkan berbagai dukungan serta bantuan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua dan keluarga yang selalu memberikan doa, dukungan materil, nasihat, dan motivasi dalam proses penyusunan Tugas Akhir.
2. Ibu Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D selaku Koordinator Program Studi Statistika FMIPA ULM atas segala ilmu dan dukungan yang telah diberikan selama perkuliahan di Program Studi Statistika FMIPA ULM.
3. Bapak Fuad Muhajirin Farid, S.Pd., M.Si dan Ibu Selvi Annisa, S.Si., M.Si selaku dosen pembimbing yang telah memberikan bimbingan dan arahan serta dukungan dalam proses penyusunan Tugas Akhir.
4. Ibu Dewi Sri Susanti, S.Si., M.Si dan Bapak Oni Soesanto, S.Si., M.Si selaku dosen penguji yang telah memberikan masukan, saran, serta bantuan dalam proses perbaikan Tugas Akhir.
5. Seluruh dosen dan staf Program Studi Statistika FMIPA ULM yang telah memberikan ilmu, nasihat, serta dukungan selama masa perkuliahan di Program Studi Statistika FMIPA ULM.
6. Teman-teman yang selalu menemani, mendukung, serta mendengarkan keluh-kesah saya selama penulisan Tugas Akhir maupun selama masa perkuliahan, khususnya teman-teman "PBI", "Anak MIA One-Two", dan "Penjoki Handal".
7. Teman-teman Program Studi Statistika FMIPA ULM khususnya Angkatan 2021 yang telah berjuang bersama selama perkuliahan.
8. Genshin Impact dan Zenless Zone Zero yang selalu membantu saya meredakan rasa stres selama penulisan Tugas Akhir.
9. Pihak-pihak lain yang berkontribusi baik secara langsung maupun tidak langsung, yang tidak dapat penulis sebutkan satu per satu.

Penulisan Tugas Akhir ini tentu masih terdapat kekurangan sehingga kritik dan saran yang membangun dari semua pihak sangat diharapkan dapat membantu dalam penyempurnaan Tugas Akhir ini. Akhir kata, penulis berharap agar hasil penelitian pada Tugas Akhir ini dapat bermanfaat bagi semua pihak.

Banjarbaru, 25 Februari 2025

Novia Ramadhani Putri Zaskia

## DAFTAR ISI

|   |     |
|---|-----|
| PERNYATAAN .....  | ii  |
| ABSTRAK.....  | iv  |
| ABSTRACT .....  | v   |
| PRAKATA.....  | vi  |
| DAFTAR ISI.....   | vii |
| DAFTAR GAMBAR.....  | ix  |
| DAFTAR TABEL.....   | x   |
| DAFTAR LAMPIRAN.....  | xi  |
| BAB I PENDAHULUAN.....  | 1   |
| 1.1 .Latar Belakang.....  | 1   |
| 1.2 .Rumusan Masalah .....  | 3   |
| 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 Diabetes.....   | 6   |
| 2.2.2 Klasifikasi.....  | 7   |
| 2.2.3 Metode <i>Ensemble</i> dan <i>Bagging</i> .....                               | 8   |
| 2.2.4 <i>Random Forest</i> .....  | 8   |
| 2.2.5 <i>Recursive Feature Elimination with Cross Validation</i> .....              | 11  |
| 2.2.6 <i>Synthetic Minority Oversampling Technique – Nominal Continuous</i> ...     | 11  |
| 2.2.7 Ukuran Ketepatan Klasifikasi .....  | 13  |
| BAB III METODE PENELITIAN.....  | 16  |
| 3.1 .Sumber Data.....   | 16  |
| 3.2 .Variabel Penelitian .....  | 16  |
| 3.3 .Prosedur Penelitian.....   | 18  |
| BAB IV HASIL DAN PEMBAHASAN .....   | 22  |
| 4.1 . <i>Pre-Processing</i> .....   | 22  |
| 4.2 .Eksplorasi Data.....   | 22  |
| 4.2.1 Deskripsi Statistik Antar Variabel Kategorik dengan Indikasi<br>Diabetes..... | 23  |
| 4.2.2 Deskripsi Statistik Antar Variabel Numerik dengan Indikasi<br>Diabetes.....   | 30  |
| 4.3 .Pembagian Data <i>Training</i> dan Data <i>Testing</i> .....                   | 32  |
| 4.4 .Pembangunan Model <i>Random Forest</i> .....                                   | 32  |
| 4.4.1 <i>Bootstrap Sampling</i> .....   | 33  |
| 4.4.2 Pembuatan Pohon Keputusan.....  | 33  |
| 4.4.3 Pengujian Model dan Evaluasi .....  | 34  |
| 4.5 .Seleksi Fitur RFECV .....  | 35  |
| 4.5.1 Iterasi Awal.....   | 35  |
| 4.5.2 Eliminasi Variabel.....   | 35  |
| 4.5.3 Evaluasi Model di Setiap Iterasi .....  | 36  |
| 4.6 .Teknik <i>Resampling</i> SMOTE-NC.....   | 36  |

|  |    |
|--|----|
| 4.6.1 Menghitung Median dari Standar Deviasi Variabel Numerik.....                   | 36 |
| 4.6.2 Menghitung Jarak Euclidean.....  | 37 |
| 4.6.3 Membuat Data Sintetis.....   | 37 |
| 4.7 .Performa Model <i>Random Forest</i> dengan Penerapan RFECV dan<br>SMOTE-NC..... | 38 |
| BAB V PENUTUP .....  | 42 |
| 5.1 .Kesimpulan .....  | 42 |
| 5.2 .Saran.....  | 42 |
| DAFTAR PUSTAKA .....   | 43 |
| LAMPIRAN.....  | 48 |
| RIWAYAT HIDUP .....  | 83 |

## DAFTAR GAMBAR

|            |   |    |
|------------|---|----|
| Gambar 2.1 | Struktur Model Random Forest .....  | 9  |
| Gambar 2.2 | Ilustrasi SMOTE.....  | 12 |
| Gambar 2.3 | Kurva ROC-AUC .....   | 15 |
| Gambar 3.1 | Alur Penelitian .....   | 21 |
| Gambar 4.1 | <i>Pie Chart</i> Proporsi Kelas Data Variabel Target.....   | 23 |
| Gambar 4.2 | Proporsi Indikasi Diabetes berdasarkan <i>HighBP, HighChol, CholCheck, dan Smoker</i> .....                 | 24 |
| Gambar 4.3 | Proporsi Indikasi Diabetes berdasarkan <i>Stroke, HeartDiseaseorAttack, PhysActivity, dan Veggies</i> ..... | 25 |
| Gambar 4.4 | Proporsi Indikasi Diabetes berdasarkan <i>HvyAlcoholConsump, AnyHealthCare, GenHlth, dan DiffWalk</i> ..... | 27 |
| Gambar 4.5 | Proporsi Indikasi Diabetes berdasarkan <i>Sex, Age, Education, dan Income</i> .....                         | 28 |
| Gambar 4.6 | Proporsi Indikasi Diabetes berdasarkan <i>Fruits dan NoDocbcCost</i> .....                                  | 30 |
| Gambar 4.7 | Sebaran <i>BMI, MentHlth, dan PhysHlth</i> berdasarkan Indikasi Diabetes.....                               | 31 |
| Gambar 4.8 | Kurva ROC Model C30 .....   | 40 |

## DAFTAR TABEL

|  |    |
|--|----|
| Tabel 2.1 Penelitian Terdahulu.....  | 4  |
| Tabel 2.2 Confusion Matrix.....  | 14 |
| Tabel 2.3 Tingkatan Nilai AUC.....   | 15 |
| Tabel 3.1 Variabel Penelitian.....   | 16 |
| Tabel 4.1 Proporsi Kelas Data Variabel Target.....                                       | 22 |
| Tabel 4.2 Pembagian Data.....  | 32 |
| Tabel 4.3 Contoh Data untuk Menghitung Jarak Euclidean.....                              | 37 |
| Tabel 4.4 Proporsi Kelas Data <i>Training</i> Sebelum dan Sesudah <i>Resampling</i> .... | 37 |
| Tabel 4.5 <i>Confusion Matrix</i> Model C30 .....  | 38 |
| Tabel 4.6 Ringkasan Model Terbaik di Masing-Masing Jenis Pembagian<br>Data .....         | 39 |

## DAFTAR LAMPIRAN

|   |    |
|---|----|
| Lampiran 1. Data Setelah Penghapusan Data Duplikat.....   | 48 |
| Lampiran 2. Kombinasi Model.....  | 49 |
| Lampiran 3. Evaluasi Kinerja Model <i>Random Forest</i> Tanpa Penerapan<br>RFECV dan SMOTE-NC .....   | 52 |
| Lampiran 4. Variabel Prediktor Optimal Terpilih di Setiap Model .....   | 55 |
| Lampiran 5. Evaluasi Kinerja Model <i>Random Forest</i> dengan Penerapan<br>RFECV dan SMOTE-NC .....  | 59 |
| Lampiran 6. <i>Syntax Python</i> untuk <i>Import Library, Data, dan</i><br><i>Pre-processing</i> .....  | 62 |
| Lampiran 7. <i>Syntax Python</i> untuk Eksplorasi Data .....  | 63 |
| Lampiran 8. <i>Syntax Python</i> untuk Pembagian Data <i>Training</i> dan Data<br><i>Testing</i> , serta <i>Dictionary</i> untuk Menyimpan Kombinasi<br>Parameter dan Semua Dataset ..... | 69 |
| Lampiran 9. <i>Syntax Python</i> untuk Model <i>Random Forest</i> Tanpa Penerapan<br>RFECV dan SMOTE-NC .....   | 70 |
| Lampiran 10. <i>Syntax Python</i> untuk RFECV .....   | 73 |
| Lampiran 11. <i>Syntax Python</i> untuk SMOTE-NC.....   | 76 |
| Lampiran 12. <i>Syntax Python</i> untuk Model <i>Random Forest</i> dengan Penerapan<br>RFECV dan SMOTE-NC .....   | 79 |
| Lampiran 13. <i>Syntax Python</i> untuk Kurva ROC Model C30 .....   | 82 |