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 FMIPA 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

Alhamdulillahi 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

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN