ANALISIS KINERJA *RANDOM FOREST* DALAM KLASIFIKASI *IMBALANCE* DATASETS MENGGUNAKAN *SMOTE* DAN *NEAR MISS*

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Sistem Informasi

disusun oleh

MUHAMMAD RIFKI AUFA

21.12.1980

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

ANALISIS KINERJA *RANDOM FOREST* DALAM KLASIFIKASI *IMBALANCE* DATASETS MENGGUNAKAN *SMOTE* DAN *NEAR MISS*

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Sistem Informasi

disusun oleh

MUHAMMAD RIFKI AUFA

21.12.1980

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

SKRIPSI

ANALISIS KINERJA RANDOM FOREST DALAM KLASIFIKASI IMBALANCE DATASETS MENGGUNAKAN SMOTE DAN NEAR MISS

yang disusun dan diajukan oleh

Muhammad Rifki Aufa

21.12.1980

telah disetujui oleh Dosen Pembimbing Skripsi

pada tanggal 04 November 2024

Dosen Pembimbing,

Yoga Pristyanto, S.Kom, M.Eng

NIK. 190302412

HALAMAN PENGESAHAN

SKRIPSI

ANALISIS KINERJA RANDOM FOREST DALAM KLASIFIKASI IMBALANCE DATASETS MENGGUNAKAN SMOTE DAN NEAR MISS

yang disusun dan diajukan oleh

Muhammad Rifki Aufa
21.12.1980
Telah dipertahankan di depan Dewan Penguji
pada tanggal 25 Juli 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Moch Farid Fauzi, S.Kom., M.Kom **NIK. 190302284**	
Wiwi Widayani, S.Kom., M.Kom **NIK. 190302272**	
Yoga Pristyanto, S.Kom., M.Eng **NIK. 190302412**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 25 Juli 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini., M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Rifki Aufa

NIM : 21.12.1980

Menyatakan bahwa Skripsi dengan judul berikut:

Analisis Kinerja *Random Forest* dalam Klasifikasi *Imbalance* datasets menggunakan SMOTE dan Near Miss

Dosen Pembimbing : Yoga Pristyanto, S.Kom., M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 25 Juli 2025

Yang Menyatakan,

Muhammad Rifki Aufa

HALAMAN PERSEMBAHAN

Syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas rahmat dan hidayah-Nya yang mempermudah penulis dalam menyelesaikan skripsi ini. Penulis juga berterima kasih kepada semua pihak yang telah memberikan dukungan, inspirasi, dan motivasi, baik secara langsung maupun tidak langsung, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Penulis mempersembahkan skripsi ini kepada:

1. Bapak Nur Azis dan Ibu Siti Khotimah selaku orang tua penulis.
2. Bapak Yoga Pristyanto, S.Kom., M.Eng, selaku dosen pembimbing penulis.
3. Rekan Slametan yang mendukung dan membantu saya dengan sepenuh hati.
4. Teman-teman rm.radongan yang membantu dan menghibur saya dari awal hingga akhirnya berada di titik ini.
5. Untuk kamu, terima kasih telah hadir dan mendukung, bahkan di saat-saat yang tidak mudah.

Semoga persembahan ini dapat mengungkapkan rasa terima kasih penulis kepada semua yang telah berperan dalam kesuksesan penyelesaian skripsi ini.

KATA PENGANTAR

Puji dan Syukur penulis panjatkan kehadirat Allah SWT, karena nikmat dan karunia-Nya penulis masih bisa menempuh pendidikan sampai saat ini, sekaligus dapat menyelesaikan skripsi yang berjudul "Analisis Kinerja *Random Forest* dalam Klasifikasi *Imbalance* datasets menggunakan SMOTE dan Near miss" ini.

Tak lupa penulis sampaikan rasa terima kasih untuk pihak-pihak yang telah berperan dalam penyusunan skripsi ini. Diantaranya adalah sebagai berikut:

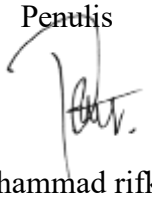
1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Ibu Prof. Dr. Kusrini., M.Kom. Selaku Dekan Fakultas Ilmu Komputer.
3. Bapak Anggit Dwi Hartanto, M.Kom. Selaku Ketua prodi Sistem Informasi.
4. Bapak Yoga Pristyanto, S.Kom., M.Eng yang telah membimbing saya dalam proses menyusun skripsi ini.
5. Seluruh keluarga saya, terutama kedua orang tua yang telah membimbing dan mendidik saya hingga sedewasa ini.
6. Teman - teman saya yang dengan setulus hati membantu penelitian ini hingga tuntas.

Masih banyak hal yang dirasa kurang dalam penyusunan proposal ini, maka dari itu penulis membuka hati dan diri sebesar-besarnya apabila ada kritik maupun saran guna melengkapi proposal ini.

Akhirnya semoga skripsi ini bisa bermanfaat bagi semua, serta menjadi acuan dan tuntunan penulis dalam membuat karya kedepan. Terima kasih.

Yogyakarta, 25 Juli 2025

Penulis



Muhammad rifki aufa

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PERSETUJUAN ii
HALAMAN PENGESAHAN iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI iv
HALAMAN PERSEMBAHAN v
KATA PENGANTAR vi
DAFTAR ISI viii
DAFTAR TABEL xii
DAFTAR GAMBAR xiii
DAFTAR LAMPIRAN xv
DAFTAR LAMBANG DAN SINGKATAN xvi
DAFTAR ISTILAH xvii
ABSTRACT xxi
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Batasan Masalah 2
1.4 Tujuan Penelitian 3
1.5 Manfaat Penelitian 3
1.6 Sistematika Penulisan 3
BAB II TINJAUAN PUSTAKA 5

2.1 Studi Literatur 5
2.2 Dasar Teori 10
2.2.1 Machine Learning 10
2.2.2 Cara Kerja Machine Learning 11
2.2.3 Klasifikasi 12
2.2.4 Imbalance Dataset 13
2.2.5 Metode Resampling 13
2.2.6 Random Forest 14
2.2.7 Model Evaluation 14
2.2.8 Confusion Matrix 14
2.2.9 Accuracy 15
2.2.10 Precision 16
2.2.11 Recall 16
2.2.12 F-1 Score 16
2.2.13 ROC AUC Score 17
2.2.14 Feature Importance pada Random Forest 17
2.2.15 Streamlit 17
BAB III METODE PENELITIAN 19
3.1 Objek Penelitian 19
3.2 Alur Penelitian 21
3.2.1 Data Collection 22
3.2.2 Data Understanding 22
3.2.3 Data Preprocessing 24
3.2.4 Handling Imbalanced 25
3.2.5 Train Model 26

3.2.6 Model Evaluation 26
3.2.7 Deploy 27
3.3 Alat dan Bahan 28
3.3.1 Hardware 28
3.3.2 Software 29
BAB IV HASIL DAN PEMBAHASAN 30
4.1 Data Collection 30
4.2 Preprocessing Data 33
4.2.1 Imputasi Nilai Kosong 33
4.2.2 Encoding Label Kategorikal 34
4.2.3 Skala dan Pemisahan Fitur 35
4.2.4 Deteksi Kolom ID Unik 36
4.3 Penanganan ketidakseimbangan Data 36
4.3.1 Implementasi SMOTE 37
4.3.2 Implementasi NearMiss 37
4.3.3 Validasi dan Perbandingan 38
4.4 Pelatihan Model 39
4.4.1 Pemilihan Random Forest sebagai Model Klasifikasi 39
4.4.2 Proses Internal pada Pelatihan 40
4.4.3 Implementasi Sistem Otomatis 40
4.4.4 Kesiapan Model untuk Evaluasi 41
4.5 Evaluasi Model 41
4.5.1 *Precision* per Dataset 41
4.5.2 *Recall* per *Dataset* 42
4.5.3 *F1- Score* Per Dataset 43

4.5.4 *Accuracy* per *Dataset* 44
4.5.5 ROC AUC Score per *Dataset* 44
4.5.6 *Confusion* Matrix 45
4.6 Prediksi Manual 45
4.6.1 Desain dan Logika Antarmuka 45
4.6.2 Proses Pengolahan Input dan Encoding 46
4.6.3 Prediksi dan Interpretasi Hasil 47
4.6.4 Fungsi Strategis Prediksi Manual 48
4.7 Deployment 48
4.7.1 Persiapan Deployment 49
4.7.2 Proses Deployment ke Streamlit Cloud 50
4.7.3 Penanganan Error dan Debugging 51
4.7.4 Hasil Deployment 52
4.7.5 Keunggulan dan Dampak Deployment 52
4.8 Klasifikasi Dataset Berdasarkan Tingkat Ketidakseimbangan 53
BAB V PENUTUP 55
5.1 Kesimpulan 55
5.2 Saran 56
REFERENSI 58
LAMPIRAN 62

DAFTAR TABEL

Tabel 2. 1Keaslian penelitian 7
Tabel 3.1 Hardware 28
Tabel 3.2 Software 29
Tabel 4.1 Precision per Dataset 41
Tabel 4.2 Recall Per Dataset 42
Tabel 4. 3F1-Score per Dataset 43
Tabel 4.4 Accuracy per Dataset 44
Tabel 4.5 ROC AUC Score per Dataset 44

DAFTAR GAMBAR

Gambar 2.1 Confusion Matrix (Sumber: Ksnugroho.medium.com, diakses 10 Mei 2025).....15
Gambar 3.1 Alur Penelitian (Sumber: Dokumen Pribadi, 2025) 21
Gambar 4.1Tampilan Koding Upload Dataset (Sumber: Dokumen Pribadi, 2025) 30
Gambar 4.2 Tampilan Upload Dataset melalui Sidebar Aplikasi (Sumber: Dokumen Pribadi, 2025) 31
Gambar 4.3 Pratinjau Dataset (Sumber: Dokumen Pribadi, 2025) 31
Gambar 4.4 Statik Deskriptif (Sumber: Dokumen Pribadi, 2025) 32
Gambar 4.5 Kode Otomatisasi Pemilihan Fitur (Sumber: Dokumen Pribadi, 2025) 32
Gambar 4.6 Pemilihan Kolom Target (Label) untuk Klasifikasi (Sumber: Dokumen Pribadi, 2025) 33
Gambar 4.7 Kode Mengganti Nilai Kosong dengan Nilai Modus (Sumber: Dokumen Pribadi, 2025) 33
Gambar 4.8 Tampilan Dataset Sebelum Imputasi Nilai Kosong (Sumber: Dokumen Pribadi, 2025) 34
Gambar 4.9 Tampilan Dataset Setelah Imputasi Nilai Kosong (Sumber: Dokumen Pribadi, 2025) 34
Gambar 4.10 Kode Label Encoding (Sumber: Dokumen Pribadi, 2025).....35
Gambar 4.11 Kode Pemisah Label X dan Y (Sumber: Dokumen Pribadi, 2025) .35
Gambar 4.12 Kode Split Data Train dan Data Testing (Sumber: Dokumen Pribadi, 2025).....36
Gambar 4.13 Contoh Kolom ID Unik yang Dihapus Secara Otomatis (Sumber: Dokumen Pribadi, 2025) 36
Gambar 4.14 Kode Penanganan Imbalance (Sumber: Dokumen Pribadi, 2025) 37
Gambar 4.15 Tampilan Pilihan Metode SMOTE pada Antarmuka Aplikasi (Sumber: Dokumen Pribadi, 2025) 37
Gambar 4.16 Kode Implementasi NearMiss (Sumber: Dokumen Pribadi, 2025) .38

Gambar 4.17 Tampilan Pilihan Metode NearMiss pada Antarmuka Aplikasi (Sumber: Dokumen Pribadi, 2025) 38
Gambar 4.18 Distribusi Kelas Setelah Proses Penyeimbangan (Sumber: Dokumen Pribadi, 2025) 38
Gambar 4.19 Kode Train Model dengan Random Forest (Sumber: Dokumen Pribadi, 2025) 39
Gambar 4.20 Kode Train Model (Sumber: Dokumen Pribadi, 2025) 41
Gambar 4.21 Kode User Interface (Sumber: Dokumen Pribadi, 2025) 46
Gambar 4.22 Antarmuka Form Input Manual yang Disesuaikan Otomatis Berdasarkan Dataset (Sumber: Dokumen Pribadi, 2025) 46
Gambar 4.23 Kode Pengolahan Input dan Encoding (Sumber: Dokumen Pribadi, 2025).....47
Gambar 4.24 Kode Prediksi Model (Sumber: Dokumen Pribadi, 2025) 47
Gambar 4.25 Hasil Prediksi Akhir dalam Format Label Asli (Sumber: Dokumen Pribadi, 2025) 48
Gambar 4.26 Struktur File requirements.txt (Sumber: Dokumen Pribadi, 2025) 50
Gambar 4.27 Struktur File Repositori GitHub yang Siap Dideploy (Sumber: Dokumen Pribadi, 2025) 50
Gambar 4. 28Konfigurasi Awal Proses Deployment di Streamlit Cloud (Sumber: Dokumen Pribadi, 2025) 51
Gambar 4. 29Tampilan Aplikasi Setelah Berhasil Dideploy di Streamlit Cloud (Sumber: Dokumen Pribadi, 2025) 52

DAFTAR LAMPIRAN

Lampiran 1. Kode Preprocessing Data.....62
Lampiran 2. Implementasi SMOTE dan NearMiss.....62
Lampiran 3. Training Model Random Forest.....62
Lampiran 4. Evaluasi Model.....62
Lampiran 5. Deploy Aplikasi Menggunakan Streamlit.....62

DAFTAR LAMBANG DAN SINGKATAN

SVM	Support Vector Machines
SMOTE	Synthetic Minority Over-sampling Technique
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
TP	True Positive – Prediksi benar untuk kelas positif
TN	True Negative – Prediksi benar untuk kelas negatif
FP	False Positive – Prediksi salah, memprediksi positif padahal negatif
FN	False Negative – Prediksi salah, memprediksi negatif padahal positif
F1-Score	Harmonik rata-rata antara precision dan recall
X	Fitur input (variabel independen)
y	Target/Label (variabel dependen)
CSV	Comma-Separated Values (format file data)
UI	User Interface (Antarmuka Pengguna)
API	Application Programming Interface
ML	Machine Learning
AI	Artificial Intelligence
TF-IDF	Term Frequency - Inverse Document Frequency
PCA	Principal Component Analysis
ID	Identifier unik pada dataset
VSCode	Visual Studio Code (editor pemrograman)
RAM	Random Access Memory
CPU	Central Processing Unit

DAFTAR ISTILAH

Artificial Intelligence (AI)	: Teknologi kecerdasan buatan yang memungkinkan mesin meniru perilaku manusia, seperti berpikir dan belajar dari data.
Machine Learning (ML)	: Cabang dari AI yang berfokus pada algoritma yang memungkinkan komputer belajar dari data tanpa diprogram secara eksplisit.
Random Forest	: Algoritma klasifikasi berbasis ensemble learning yang menggunakan banyak pohon keputusan untuk meningkatkan akurasi dan stabilitas model.
Decision Tree	: Model klasifikasi berbentuk struktur pohon yang membuat keputusan berdasarkan atribut dari data.
Klasifikasi	: Proses dalam machine learning untuk memetakan data ke dalam kelas atau kategori tertentu.
Imbalanced Dataset	: Dataset dengan distribusi kelas yang tidak seimbang, sehingga model cenderung bias terhadap kelas mayoritas.
SMOTE	: Teknik oversampling yang menghasilkan data sintetis baru pada kelas minoritas untuk menyeimbangkan distribusi kelas.
NearMiss	: Teknik undersampling yang menghapus sebagian data dari kelas mayoritas berdasarkan kedekatan jarak dengan kelas minoritas.
Oversampling	: Teknik penyeimbangan data dengan menambahkan salinan atau data baru pada kelas minoritas.
Undersampling	: Teknik penyeimbangan data dengan mengurangi jumlah data dari kelas mayoritas.
Confusion Matrix	: Matriks evaluasi yang menampilkan hasil klasifikasi model dalam bentuk jumlah prediksi benar dan salah.

True Positive (TP)	: Prediksi benar terhadap data yang memang termasuk kelas positif.
True Negative (TN)	: Prediksi benar terhadap data yang memang termasuk kelas negatif.
False Positive (FP)	: Prediksi salah, di mana model memprediksi positif padahal data sebenarnya negatif.
False Negative (FN)	: Prediksi salah, di mana model memprediksi negatif padahal data sebenarnya positif.
Accuracy	: Rasio antara jumlah prediksi benar dengan total keseluruhan data.
Precision	: Rasio prediksi benar terhadap kelas positif dibandingkan dengan semua prediksi kelas positif.
Recall (Sensitivity)	: Rasio prediksi benar terhadap kelas positif dibandingkan dengan jumlah aktual kelas positif.
F1-Score	: Rata-rata harmonik dari precision dan recall yang digunakan untuk mengukur keseimbangan antara keduanya.
ROC	: Kurva yang menunjukkan performa model klasifikasi pada berbagai ambang batas.
AUC (Area Under Curve)	: Nilai area di bawah kurva ROC yang menggambarkan kemampuan model membedakan antara kelas.
Label Encoding	: Proses mengubah data kategorikal menjadi angka berdasarkan urutan nilai.
One-Hot Encoding	: Teknik encoding yang mengubah data kategorikal menjadi representasi biner (0 dan 1).
Feature (Fitur)	: Atribut atau kolom dalam dataset yang digunakan sebagai input untuk pelatihan model.
Target/Label	: Atribut dalam dataset yang ingin diprediksi oleh model (output).

Train-Test Split	: Proses membagi dataset menjadi dua bagian: data latih dan data uji.
Preprocessing	: Tahapan pembersihan dan persiapan data sebelum digunakan dalam pelatihan model.
Streamlit	: Framework Python untuk membangun aplikasi web secara interaktif dan cepat, khususnya untuk data science.
CSV	: Format file data berbentuk tabel dengan nilai yang dipisahkan oleh koma.
Cross Validation	: Teknik evaluasi model dengan cara membagi dataset ke dalam beberapa subset untuk pengujian silang.
Stratified Sampling	: Teknik pengambilan sampel yang mempertahankan distribusi kelas asli dalam dataset.
Dataset	: Kumpulan data yang digunakan untuk membangun dan menguji model machine learning

INTISARI

Ketidakseimbangan dalam dataset menjadi tantangan dalam klasifikasi, terutama dalam dataset di bidang- bidang penting seperti Kesehatan, keuangan, dan pemeliharaan mesin. Penelitian ini akan membandingkan kinerja metode Synthetic Minority Over-sampling Technique (SMOTE) dan NearMiss dalam menangani ketidakseimbangan data dengan menggunakan algoritma Random Forest. Sepuluh dataset yang digunakan dalam penelitian ini mencakup Credit card Fraud, Predictive Maintenance, Diabetes, Spam SMS, Customer Churn, Breast Cancer, Spam Email, dan Employee Performance.

Hasil evaluasi di dalam penelitian ini menunjukan bahwa SMOTE memberikan hasil yang baik dalam empat dari sepuluh dataset berdasarkan metrik *precision, recall, F1-score*, *ROC AUC Score* dan *Confusion Matrix* apalagi dalam kasus dataset di mana kelas minoritas memiliki jumlah data yang sangat sedikit. Sebaliknya, NearMiss menunjukan performa yang lebih baik dalam satu dataset, sedangkan lima dataset lainnya memiliki hasil yang seimbang antara kedua metode.

Secara keseluruhan, dalam penelitian ini SMOTE cenderung lebih efektif dalam meningkatkan recall kelas minoritas, sedangkan NearMiss dapat mengurangi bias dalam model dengan menyeimbangkan jumlah sampel dataset. Hasil dari penelitian ini menjadi temuan dan memberikan wawasan tentang bagaimana pemilihan metode penyeimbangan dataset dapat mempengaruhi performa model klasifikasi dalam berbagai scenario dataset.

Kata kunci: klasifikasi, ketidakseimbangan data, SMOTE, NearMiss, Random Forest.

ABSTRACT

Imbalance in datasets is a challenge in classification, especially in datasets in important areas such as Health, finance, and machine maintenance. This research will compare the performance of the Synthetic Minority Over-sampling Technique (SMOTE) and NearMiss methods in handling data imbalance using the Random Forest algorithm. The ten datasets used in this research include Credit card Fraud, Predictive Maintenance, Diabetes, SMS Spam, Customer Churn, Breast Cancer, GiveMeSomeCredit, Email Spam, and Employee Performance.

The evaluation results in this study show that SMOTE provides good results in four out of ten datasets based on precision, recall, f1-score metrics, ROC AUC Score and Confusion Matrix especially in the case of datasets where the minority class has a very small amount of data. In contrast, NearMiss shows better performance in one dataset, while the other five datasets have balanced results between the two methods.

Overall, in this study SMOTE tends to be more effective in increasing minority class recall, while NearMiss can reduce bias in the model by balancing the number of dataset samples. The results of this research are findings and provide insight into how the choice of dataset balancing method can affect the performance of classification models in various dataset scenarios.

***Keywords**: classification, data imbalance, SMOTE, Nearmiss, Random Forest.*