

TUGAS AKHIR

**ANALISIS PENGARUH PENERAPAN SMOTE PADA
AKURASI RANDOM FOREST DAN C4.5 DALAM
KLASIFIKASI DATA TIDAK SEIMBANG**



FANISA NIMASTITI

20110004

**PROGRAM STUDI SAINS DATA
FAKULTAS INFORMATIKA
INSTITUT TEKNOLOGI TELKOM PURWOKERTO
2024**

TUGAS AKHIR

**ANALISIS PENGARUH PENERAPAN SMOTE PADA
AKURASI RANDOM FOREST DAN C4.5 DALAM
KLASIFIKASI DATA TIDAK SEIMBANG**

***ANALYSIS OF THE EFFECT OF APPLYING SMOTE
ON THE ACCURACY OF RANDOM FOREST AND
C4.5 IN IMBALANCED DATA CLASSIFICATION***

Disusun Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Sains Data



FANISA NIMASTITI

20110004

**PROGRAM STUDI SAINS DATA
FAKULTAS INFORMATIKA
INSTITUT TEKNOLOGI TELKOM PURWOKERTO**

2024

LEMBAR PERSETUJUAN PEMBIMBING

**ANALISIS PENGARUH PENERAPAN SMOTE PADA
AKURASI RANDOM FOREST DAN C4.5 DALAM
KLASIFIKASI DATA TIDAK SEIMBANG**

***ANALYSIS OF THE EFFECT OF APPLYING SMOTE
ON THE ACCURACY OF RANDOM FOREST AND
C4.5 IN IMBALANCED DATA CLASSIFICATION***

Dipersiapkan dan Disusun oleh

FANISA NIMASTITI

20110004

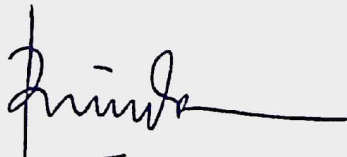
Fakultas Informatika

Institut Teknologi Telkom Purwokerto


Pada Tanggal : 18 Januari 2024

Pembimbing Utama,

Pembimbing Pendamping



(Rima Dias Ramadhani, S.Kom., M.Kom)
NIDN.0602039301



(Atika Ratna Dewi, S.Si, M.Sc)
NIDN.0615128703

**ANALISIS PENGARUH PENERAPAN SMOTE PADA
AKURASI RANDOM FOREST DAN C4.5 DALAM
KLASIFIKASI DATA TIDAK SEIMBANG**

***ANALYSIS OF THE EFFECT OF APPLYING SMOTE
ON THE ACCURACY OF RANDOM FOREST AND
C4.5 IN IMBALANCED DATA CLASSIFICATION***


Disusun oleh

FANISA NIMASTITI
20110004


Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tugas

Akhir Pada Hari Senin, Tanggal 29 Januari 2024


Penguji I,


(Siti Khamsah S.Kom.,
M.Cs)
NIDN. 0517108101

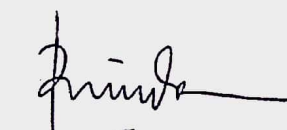
Penguji II,


(Sudianto, S.Pd., M.Kom.)
NIDN.0605049301

Penguji III,


(Aminatus Sa'adah, S.Si.,
M.Si)
NIDN.0610079602

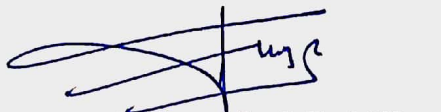
Pembimbing Utama,


(Riina Dias Ramadhani,
S.Kom., M.Kom)
NIDN. 0602039301

Pembimbing Pendamping


(Atika Ratna Dewi, S.Si.,
M.Sc)
NIDN. 0602039301

Dekan,


Auliya Burhanuddin, S.Si., M.Kom
NIK. 19820008

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertandatangan di bawah ini,

Nama Mahasiswa : Fanisa Nimastiti

NIM : 20110004

Program Studi : S1 Sains Data

Menyatakan bahwa Tugas Akhir dengan judul berikut:

ANALISIS PENGARUH PENERAPAN SMOTE PADA AKURASI RANDOM FOREST DAN C4.5 DALAM KLASIFIKASI DATA TIDAK SEIMBANG

Dosen Pembimbing Utama : Rima Dias Ramadhani, S.Kom., M.Kom

Dosen Pembimbing Pendamping : Atika Ratna Dewi, S.Si, M.Sc

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Institut Teknologi Telkom Purwokerto maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan, dan penelitian Saya Sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim 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 tanggungjawab Saya, bukan tanggungjawab Institut Teknologi Telkom Purwokerto.
5. Pernyataan ini Saya buat dengan sesungguhnya, apabila dikemudian 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.

Purwokerto, 1 Februari 2024,

Yang Menyatakan,



(Fanisa Nimastiti)

KATA PENGANTAR

Puji syukur kepada Allah SWT berkat Rahmat, Hidayah, dan Karunia-Nya kepada kita semua sehingga kami dapat menyelesaikan proposal skripsi dengan judul " Analisis Pengaruh Penerapan Smote Pada Akurasi Random Forest Dan C4.5 Dalam Klasifikasi Data Tidak Seimbang ". Laporan skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana pada program Strata-1 di Jurusan Sains Data, Fakultas Informatika. Penulis menyadari dalam penyusunan skripsi ini tidak akan selesai tanpa bantuan dari berbagai pihak. Karena itu pada kesempatan ini kami ingin mengucapkan terima kasih kepada:

- 1.) Dr. Tenia Wahyuningrum, S.Kom.,M.T selaku Rektor Institut Teknologi Telkom Purwokerto.
- 2.) Auliya Burhanuddin, S.Si.,M.Kom, selaku Dekan Fakultas Informatika
- 3.) Siti Khomsah, S.Kom, M.Cs. selaku Kaprodi S1 Sains Data Institut Teknologi Telkom Purwokerto Institut Teknologi Telkom Purwokerto
- 4.) Rima Dias Ramadhani, S.Kom., M.Kom. dan Atika Ratna Dewi, S.Si, M.Sc selaku Dosen Pembimbing atas bimbingan, saran, dan motivasi yang diberikan.
- 5.) Orang tua, atas doa, bimbingan, serta kasih sayang yang selalu tercurah selama ini.
- 6.) Kepada teman-teman terdekat saya Niesya Ayunda Febriyanti, Elisabeth Angeline, dan Nadea Putri Nur Fauzi yang telah membantu saya menyelesaikan skripsi ini

Purwokerto, 01 Februari 2024



Fanisa Nimastiti

DAFTAR ISI

HALAMAN PERSETUJUAN PEMBIMBING	iii
HALAMAN PENGESAHAN	iv
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR	xi
ABSTRAK.....	xii
ABSTRACT.....	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	2
1.3 Pertanyaan Penelitian.....	2
1.4 Batasan Masalah	3
1.5 Tujuan Penelitian	3
1.6 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA	4
2.1. Kajian Pustaka	4
2.2. Landasan Teori.....	13
2.2.1. <i>Imbalanced Dataset</i>	13
2.2.2. Pendekatan Level Data.....	13
2.2.3. <i>Oversampling</i>	14
2.2.4. <i>Imbalance Ratio</i>	16

2.2.5. Pendekatan Level Data.....	17
2.2.6. <i>Robust Scaller</i>	17
2.2.7. <i>Random Forest</i>	17
2.2.8. Algoritma <i>C4.5</i>	18
2.2.9. Metrik Evaluasi.....	20
BAB III METODOLOGI PENELITIAN	22
3.1. Subjek dan Object Penelitian	22
3.2. Alat dan Bahan.....	22
3.3. Alur Penelitian	22
3.4. Merumusan Masalah.....	22
3.5. Studi Literatur	23
3.6. Pengumpulan Data	23
3.6.1. <i>Haberman's Survival</i>	23
3.6.2. Indian Liver Patient Records.....	24
3.6.3. Adult	25
3.6.4. Credit Risk	26
3.7. Eksplorasi Data	26
3.8. <i>Data Preprocessing</i>	27
3.9. Pengujian Model	28
3.10. Evaluasi Data	29
BAB IV HASIL DAN PEMBAHASAN	30
4.1. Hasil...	30
4.1.1. Eksplorasi Data	30
4.1.1.1. Eksplorasi Data 1	30
4.1.1.2. Eksplorasi Data 2	33

4.1.1.3.	Eksplorasi Data 3	36
4.1.1.4.	Eksplorasi Data 4	39
4.1.2.	<i>Data Preprocessing</i>	42
4.1.2.1.	Penghapusan Data	43
4.1.2.2.	Imputasi Data	43
4.1.2.3.	<i>Label Encoder</i>	45
4.1.2.4.	Standarisasi	45
4.1.2.5.	<i>Oversampling</i>	46
4.1.2.5.1.	Hasil Penerapan SMOTE Data 1	48
4.1.2.5.2.	Hasil Penerapan SMOTE Data 2.....	49
4.1.2.5.3.	Hasil Penerapan SMOTE Data 3.....	51
4.1.2.5.4.	Hasil Penerapan SMOTE Data 4.....	55
4.1.3.	Pengujian Model	58
4.2.	Pembahasan.....	60
BAB V KESIMPULAN DAN SARAN		62
5.1	Kesimpulan	62
5.2	Saran	62
DAFTAR PUSTAKA		63
LAMPIRAN.....		67

DAFTAR TABEL

Tabel 2.1. Penelitian Terdahulu	7
Tabel 2. 2 Kategori Tingkat Ketidakseimbangan Data.....	16
Tabel 2. 3 Confusion Matriks	21
Tabel 3.1 Dataset Haberman’s Suvival.....	24
Tabel 4. 1 Eksplorasi Data 1	30
Tabel 4. 2 Eksplorasi Data 2	33
Tabel 4. 3 Missing Value Data 2	33
Tabel 4. 4 Eksplorasi Data 3	36
Tabel 4. 5 Missing Value Data 3	37
Tabel 4. 6 Eksplorasi Data 4	40
Tabel 4. 7 Missing Value Data 4.....	41
Tabel 4. 8 Hasil Label Encoder.....	45
Tabel 4.9 Kasus SMOTE	46
Tabel 4. 10 Hasil Pengujian Model.....	59

DAFTAR GAMBAR

Gambar 2. 1 Alur Algoritma Random Forest	18
Gambar 2. 2 Alur Algoritma C4.5	19
Gambar 3.1 Alur Penelitian	22
Gambar 3. 2 Alur Pengujian Algoritma	28
Gambar 4. 1 Distribusi Kelas Data 1	31
Gambar 4. 2 Boxplot Data 1	32
Gambar 4. 3 Distribusi Data 1	32
Gambar 4. 4 Distribusi Kelas Data 2	34
Gambar 4. 5 Boxplot Data 2	35
Gambar 4. 6 Distribusi Data 2	36
Gambar 4. 7 Distribusi Kelas Data 3	38
Gambar 4. 8 Boxplot Data 3	38
Gambar 4. 9 Distribusi Data 3	39
Gambar 4. 10 Distribusi Kelas Data 1	40
Gambar 4. 11 Pengecekan Data 4	41
Gambar 4. 12 Boxplot Data 4	42
Gambar 4. 13 Distribusi Data 4	42
Gambar 4. 14 Hasil Robust Scaller	46
Gambar 4. 15 Persebaran Data 1	49
Gambar 4. 16 Persebaran Data 2	51
Gambar 4. 17 Persebaran Variabel fnlwght dan Age	53
Gambar 4. 18 Diagram Batang Variabel Race	53
Gambar 4. 19 Diagram Batang Variabel workclass	54
Gambar 4. 20 Diagram Batang Variabel marital.status	54
Gambar 4. 21 Persebaran Variabel person_income dan loan_amnt	56
Gambar 4. 22 Diagram Batang Variabel loan_grade	57
Gambar 4. 23 Diagram Batang Person_home_ownership	57
Gambar 4. 24 Diagram Batang loan_intent	58