

**KOMBINASI *OVERSAMPLING* DAN
UNDERSAMPLING DALAM MENANGANI *CLASS*
UNBALANCED DAN *OVERLAPPING* PADA
KLASIFIKASI DATA *BANK MARKETING***

TESIS

oleh

ANAK AGUNG GDE WAHYU SUKMA ERLANGGA

NIM 2229101028



**PROGRAM STUDI ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA
SINGARAJA**

2024

KOMBINASI *OVERSAMPLING* DAN *UNDERSAMPLING*
DALAM MENANGANI *CLASS UNBALANCED* DAN
OVERLAPPING PADA KLASIFIKASI DATA *BANK*
MARKETING

TESIS

oleh

ANAK AGUNG GDE WAHYU SUKMA ERLANGGA

NIM 2229101028



PROGRAM STUDI ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA
SINGARAJA

2024

KOMBINASI *OVERSAMPLING* DAN *UNDERSAMPLING* DALAM
MENANGANI *CLASS UNBALANCED* DAN *OVERLAPPING* PADA
KLASIFIKASI DATA *BANK MARKETING*

TESIS

Diajukan kepada

Universitas Pendidikan Genesha
untuk Memenuhi Sebagian Persyaratan
Memperoleh Gelar Magister Komputer
Program Studi Ilmu Komputer

oleh

ANAK AGUNG GDE WAHYU SUKMA ERLANGGA
NIM 2229101028



PROGRAM STUDI ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS PENDIDIKAN GANESHA
SINGARAJA

2024

LEMBAR PERSETUJUAN

Tesis oleh Anak Agung Gde Wahyu Sukma Erlangga ini telah diperiksa dan disetujui untuk mengikuti Ujian Tesis

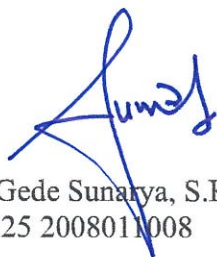
Singaraja, 4 Januari 2024

Pembimbing I



Dr. I Gede Aris Gunadi, S.Si., M.Kom.
NIP 197703182008121004

Pembimbing II




Dr. I Made Gede Sunarya, S.Kom., M.Cs.
NIP 19830725 2008011008

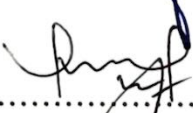
LEMBAR PERSETUJUAN TIM PENGUJI


Tesis oleh Anak Agung Gde Wahyu Sukma Erlangga ini telah dipertahankan di depan tim penguji dan dinyatakan diterima sebagai salah satu persyaratan untuk memperoleh gelar Magister Komputer di Program Studi Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

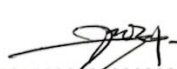
Disetujui pada tanggal: 30 Januari 2024

oleh
Tim Penguji

.....,  Ketua (Dr. I Made Gede Sunarya, S.Kom., M.Cs.)
NIP 198307252008011008

.....,  Anggota (Dr. I Gede Aris Gunadi, S.Si., M.Kom.)
NIP 197703182008121004

.....,  Anggota (Dr. I Nyoman Sukajaya, M.T.)
NIP 196711151993031001

.....,  Anggota (Dr. Gede Suweken, M.Sc.)
NIP 196111111987021001

Mengetahui Direktur
Program Pascasarjana Undiksha,



Prof. Dr. I Nyoman Jampel, M.Pd.
NIP 195910101986031003

LEMBAR PENYATAAN

Saya menyatakan dengan sesungguhnya bahwa tesis yang saya susun sebagai syarat untuk memperoleh gelar Magister Komputer dari Program Pascasarjana Universitas Pendidikan Ganesha seluruhnya merupakan hasil karya saya sendiri. Bagian-bagian tertentu dalam penulisan tesis yang saya kutip dari hasil karya orang lain telah dituliskan sumbernya secara jelas dan sesuai dengan norma, kaidah, serta etika akademis.

Apabila di kemudian hari ditemukan seluruh atau sebagian tesis ini bukan hasil karya saya sendiri atau adanya plagiat dalam bagian-bagian tertentu, saya bersedia menerima sanksi pencabutan gelar akademik yang saya sandang dan sanksi-sanksi lainnya sesuai dengan peraturan perundang-undangan yang berlaku di wilayah Negara Kesatuan Republik Indonesia.

Singaraja, 30 Januari 2024
Yang memberi pernyataan,



(Anak Agung Gde Wahyu Sukma Erlangga)

PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas anugrah-Nya, sehingga tesis yang berjudul: “Kombinasi Oversampling dan Undersampling dalam Menangani *Class Unbalanced* dan *Overlapping* pada Klasifikasi Data *Bank Marketing*”, dapat diselesaikan sesuai dengan yang direncanakan.

Tesis ini ditulis untuk memenuhi salah satu persyaratan untuk memperoleh gelar Magister Komputer Program Pascasarjana Universitas Pendidikan Ganesha pada Program Studi Ilmu Komputer. terselesaikannya tesis ini telah banyak memperoleh uluran tangan dari berbagai pihak. Untuk itu, ijin penulis menyampaikan terima kasih dan penghargaan kepada pihak-pihak berikut.

1. Bapak Dr. I Gede Aris Gunadi, S.Si., M.Kom., sebagai pembimbing I yang telah dengan sabar membimbing, mengarahkan, dan memberikan motivasi yang demikian bermakna, sehingga penulis mampu melewati berbagai hambatan dalam perjalanan studi dan penyelesaian tesis ini;
2. Bapak Dr. I Made Gede Sunarya, S.Kom., M.Cs., sebagai pembimbing II, yang dengan gaya dan pola komunikasi yang khas, telah melecut semangat, motivasi, dan harapan penulis selama penelitian dan penulisan naskah laporan tesis ini, sehingga tesis ini dapat terwujud dengan baik sesuai harapan;
3. Bapak Dr. I Nyoman Sukajaya, M.T. dan Bapak Dr. Gede Suweken, M.Sc., sebagai penguji yang telah banyak memberikan masukan-masukan yang bermanfaat untuk penyempurnaan tesis ini;
4. Koordinator Program Studi Ilmu Komputer dan staf dosen pengajar yang telah banyak membantu dan memotivasi penulis selama penyusunan tesis ini;
5. Direktur Program Pascasarjana Undiksha dan staf, yang telah banyak membantu selama penulis menyelesaikan tesis ini;
6. Rektor Universitas Pendidikan Ganesha, yang telah memberikan bantuan secara moral dan memfasilitasi berbagai kepentingan penulis dalam menyelesaikan tesis ini;
7. Rekan-rekan seangkatan di Program Studi Ilmu Komputer yang dengan karakternya masing-masing telah banyak berkontribusi membentuk kedirian penulis selama menjalani studi dan penyelesaian tesis ini;
8. Bapak Anak Agung Gede Sayang Wirawan dan Ibu Anak Agung Raka Sriani selaku orang tua penulis, yang telah banyak membantu secara material dan moral selama penyelesaian tesis ini.
9. Seluruh pihak yang telah membantu dan mendukung penulisan tesis ini, yang tidak bisa penulis sebutkan satu persatu.

Semoga semua bantuan yang telah mereka berikan dalam menyelesaikan studi ini, mereka diberkati imbalan yang sepadan oleh Tuhan Yang Maha Esa, kesehatan, dan keharmonian dalam menjalani kehidupan.

Penulis menyadari bahwa tesis ini belum sempurna. Namun, kehadirannya dalam konstelasi masyarakat akademis akan menambah perbendaharaan ilmu dalam perkembangan ilmu pengetahuan. Semoga tesis ini bermanfaat bagi

masyarakat akademis, terutama mereka yang menyatakan diri bernaung di bawah
kebesaran panji-panji pendidikan.

Singaraja, 30 Januari 2024
Penulis



DAFTAR ISI

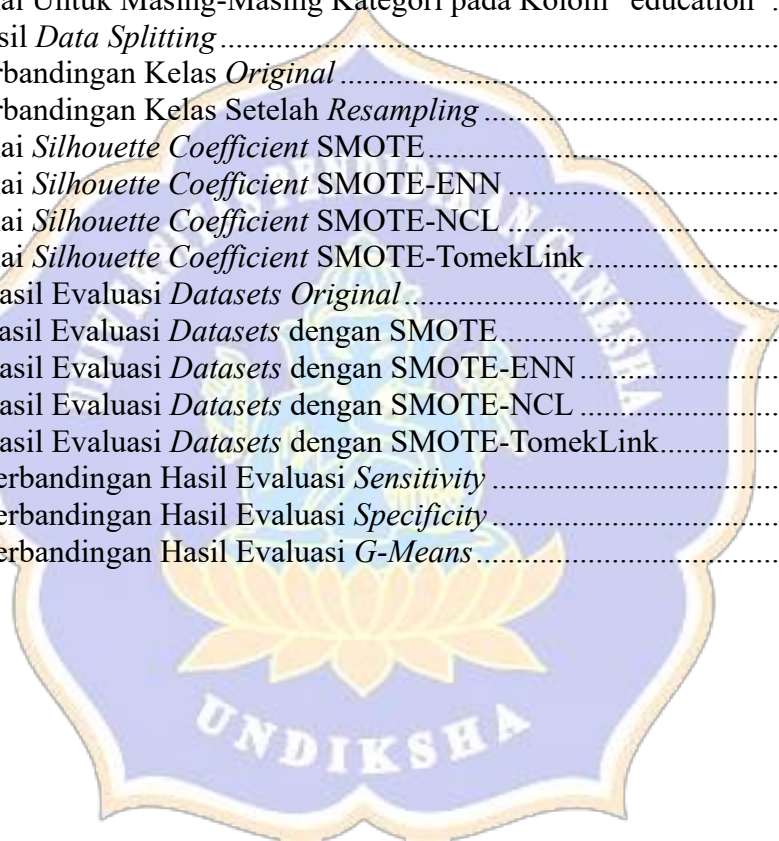
LEMBAR PERSETUJUAN	i
LEMBAR PERSETUJUAN TIM PENGUJI.....	ii
LEMBAR PENYATAAN	iii
PRAKATA.....	iv
ABSTRAK	vi
<i>ABSTRACT</i>	vii
DAFTAR ISI.....	viii
DAFTAR TABEL	x
DAFTAR GAMBAR.....	xi
DAFTAR RUMUS.....	xii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah.....	4
1.3 Batasan Masalah.....	5
1.4 Rumusan Masalah	5
1.5 Tujuan Penelitian.....	6
1.6 Manfaat Penelitian	6
1.6.1 Manfaat Teoritik.....	6
1.6.2 Manfaat Praktis	7
BAB II KAJIAN PUSTAKA	8
2.1 <i>Machine Learning</i>	8
2.1.1 <i>Logistic Regression</i>	9
2.2 <i>Class Imbalanced</i>	10
2.3 <i>Oversampling</i>	12
2.3.1 <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	12
2.4 <i>Undersampling</i>	13
2.4.1 <i>Edited Nearest Neighbor (ENN)</i>	14
2.4.2 <i>Neighborhood Cleaning Rule (NCL)</i>	14
2.4.3 <i>Tomek Link</i>	15
2.5 Kombinasi <i>Oversampling</i> dan <i>Undersampling</i>	15
2.6 Matriks Evaluasi	16
2.7 Kajian Hasil Penelitian yang Relevan.....	18

BAB III METODE PENELITIAN	22
3.1 Metode Penelitian.....	22
3.1.1 <i>Data Collection</i>	23
3.1.2 <i>Data Preprocessing</i>	25
3.1.3 <i>Data Splitting</i>	25
3.1.4 <i>Data Resampling</i>	26
3.1.5 <i>Modeling</i>	29
3.1.6 <i>Evaluating Model</i>	29
BAB IV HASIL DAN PEMBAHASAN.....	31
4.1 Data Collection.....	31
4.2 Data Preprocessing.....	32
4.2.1 Missing Value Handling.....	33
4.2.2 Encoding Data Kategorikal.....	34
4.2.3 Memisahkan Fitur dan Target	37
4.2.4 Standarisasi Data.....	39
4.3 Data Splitting	40
4.4 <i>Data Resampling</i>	40
4.5 Penguujian Hasil <i>Resampling</i> dengan <i>Silhouette Coefficient</i>	43
4.6 Klasifikasi dan Evaluasi Model	46
4.7 Pembahasan.....	50
BAB V PENUTUP	57
5.1 Simpulan	57
5.2 Saran.....	58
DAFTAR PUSTAKA.....	59
LAMPIRAN.....	65



DAFTAR TABEL

Tabel 2.1 Kajian Penelitian yang Relevan	18
Tabel 3.1 Deskripsi <i>Bank Marketing Datasets</i>	23
Tabel 3.2 Deskripsi <i>Credit Card Fraud Datasets</i>	24
Tabel 3.3 Deskripsi <i>Cerebral Stroke Datasets</i>	24
Tabel 3.4 Contoh Data.....	26
Tabel 3.5 Jarak <i>Euclidean</i> Terhadap D1	26
Tabel 3.6 Jarak <i>Euclidean</i> Antar Data	27
Tabel 3.7 Tetangga Terdekat Masing-Masing Data.....	28
Tabel 4.1 Keterangan Masing-Masing <i>Datasets</i>	31
Tabel 4.2 Nilai Untuk Masing-Masing Kategori pada Kolom “education”	36
Tabel 4.3 Hasil <i>Data Splitting</i>	40
Tabel 4.4 Perbandingan Kelas <i>Original</i>	41
Tabel 4.5 Perbandingan Kelas Setelah <i>Resampling</i>	41
Tabel 4.6 Nilai <i>Silhouette Coefficient</i> SMOTE.....	43
Tabel 4.7 Nilai <i>Silhouette Coefficient</i> SMOTE-ENN	44
Tabel 4.8 Nilai <i>Silhouette Coefficient</i> SMOTE-NCL	44
Tabel 4.9 Nilai <i>Silhouette Coefficient</i> SMOTE-TomekLink	45
Tabel 4.10 Hasil Evaluasi <i>Datasets Original</i>	46
Tabel 4.11 Hasil Evaluasi <i>Datasets</i> dengan SMOTE	47
Tabel 4.12 Hasil Evaluasi <i>Datasets</i> dengan SMOTE-ENN	48
Tabel 4.13 Hasil Evaluasi <i>Datasets</i> dengan SMOTE-NCL	49
Tabel 4.14 Hasil Evaluasi <i>Datasets</i> dengan SMOTE-TomekLink.....	49
Tabel 4.15 Perbandingan Hasil Evaluasi <i>Sensitivity</i>	50
Tabel 4.16 Perbandingan Hasil Evaluasi <i>Specificity</i>	53
Tabel 4.17 Perbandingan Hasil Evaluasi <i>G-Means</i>	54



DAFTAR GAMBAR

Gambar 2.1 <i>Confusion Matrix</i>	16
Gambar 3.1 Metode Penelitian.....	22
Gambar 4.1 <i>Datasets Bank Marketing</i>	32
Gambar 4.2 <i>Datasets Credit Card Fraud</i>	32
Gambar 4.3 <i>Dataset Cerebral Stroke</i>	32
Gambar 4.4 Hasil Imputasi <i>Missing Value Datasets Bank Marketing</i>	33
Gambar 4.5 Hasil Imputasi <i>Missing Value Datasets Cerebral Stroke</i>	34
Gambar 4.6 Hasil <i>One-Hot Encoding Datasets Bank Marketing</i>	35
Gambar 4.7 Hasil <i>Ordinal Encoding Datasets Bank Marketing</i>	36
Gambar 4.8 Hasil <i>One-Hot Encoding Datasets Cerebral Stroke</i>	37
Gambar 4.9 Hasil Pemisahan Fitur dan Target <i>Datasets Bank Marketing</i>	38
Gambar 4.10 Hasil Pemisahan Fitur dan Target <i>Datasets Cerebral Stroke</i>	38
Gambar 4.11 Hasil Pemisahan Fitur dan Target <i>Datasets Credit Card Fraud</i>	38
Gambar 4.12 Hasil Standarisasi Fitur <i>Datasets Bank Marketing</i>	39
Gambar 4.13 Hasil Standarisasi Fitur <i>Datasets Cerebral Stroke</i>	39
Gambar 4.14 Hasil Standarisasi Fitur <i>Datasets Credit Card Fraud</i>	40



DAFTAR RUMUS

(2.1) Rumus <i>Logistic Regression</i>	9
(2.2) Rumus <i>Synthetic Minority Oversampling (SMOTE)</i>	12
(2.3) Rumus <i>Sensitivity</i>	17
(2.4) Rumus <i>Specificity</i>	17
(2.5) Rumus <i>G-Means</i>	18



DAFTAR LAMPIRAN

Lampiran 1. Datasets Bank Marketing.....	65
Lampiran 2. Datasets Credit Card Fraud.....	65
Lampiran 3. Datasets Cerebral Stroke	65

