

DAFTAR ISI

HALAMAN JUDUL.....	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PERNYATAAN KEASLIAN TESIS	v
HALAMAN PERSEMBAHAN	vi
HALAMAN MOTTO	vii
KATA PENGANTAR	viii
DAFTAR ISI.....	x
DAFTAR TABEL.....	xii
DAFTAR GAMBAR	xiii
DAFTAR ISTILAH	xiv
INTISARI.....	xx
ABSTRACT.....	xxi
BAB I PENDAHULUAN	
1.1. Latar Belakang Masalah	1
1.2. Rumusan Masalah	5
1.3. Batasan Masalah	6
1.4. Tujuan Penelitian.....	6
1.5. Manfaat Penelitian.....	7
BAB II TINJAUAN PUSTAKA	
2.1. Tinjauan Pustaka	8
2.2. Keaslian Penelitian	13
2.3. Landasan Teori	16
2.3.1 Class Imbalance pada Machine Learning	16
2.3.2 SMOTE (<i>Synthetic Minority Over-sampling Technique</i>)	17
2.3.3 Borderline-SMOTE.....	18
2.3.4 ADASYN (<i>Adaptive Synthetic Sampling</i>)	20
2.3.5 SMOTE-ENN (<i>Synthetic Minority Over-sampling Technique - Edited Nearest Neighbor</i>)	21
2.3.6 XGBoost (<i>Extreme Gradient Boosting</i>).....	23

2.3.7 LightGBM (Light Gradient Boosting Machine)	24
2.3.8 Metrik Evaluasi untuk Data Imbalance.....	26
BAB III METODE PENELITIAN	
3.1. Jenis, Sifat, dan Pendekatan Penelitian	28
3.2. Metode Pengumpulan Data	29
3.3. Metode Analisis Data	30
3.4. Alur Penelitian.....	31
3.4.1 Preprocessing Data.....	35
3.4.2 Pelatihan dan Evaluasi Model.....	36
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	
4.1. Hasil Analisis Komparatif Performa Empat Teknik Oversampling.....	45
4.1.1 Hasil pada Dataset Bank Marketing (IR 7,55:1).....	47
4.1.2 Hasil pada Dataset Glass Identification (IR 15,46:1).....	49
4.1.3 Hasil pada Dataset Musk Version I (IR 1,30:1).....	50
4.2. Hasil Performa Kombinasi Teknik Oversampling dan Classifier	52
4.3. Hasil Analisis Pola Interaksi Teknik Oversampling dengan Classifier Boosting.....	55
4.4. Pembahasan	58
4.4.1 Pembahasan Hasil pada Dataset Bank Marketing	58
4.4.2 Pembahasan Hasil pada Dataset Glass Identification.....	59
4.4.3 Pembahasan Hasil pada Dataset Musk Version I	60
4.4.4 Pembahasan Hasil Kombinasi Teknik Oversampling dan Classifier 61	
4.4.5 Pembahasan Pola Interaksi Teknik Oversampling dengan Classifier Boosting	62
BAB V PENUTUP	
5.1. Kesimpulan.....	66
5.2. Saran	68
DAFTAR PUSTAKA	
LAMPIRAN	

DAFTAR TABEL

Tabel 2.1 Matriks literatur review dan posisi penelitian Analisis Komparatif Teknik Oversampling Terhadap Performa XGBoost dan LightGBM pada Klasifikasi Data Imbalance Multi-Dataset.....	13
Tabel 3.1 Dataset Penelitian.....	29
Tabel 3.2 Desain Eksperimen Komparatif	36
Tabel 4.1 Pseudocode SMOTE (Synthetic Minority Over-sampling Technique)	38
Tabel 4.2 Pseudocode Borderline-SMOTE.....	39
Tabel 4.3 Pseudocode ADASYN (Adaptive Synthetic Sampling)	40
Tabel 4. 4 Pseudocode SMOTE-ENN (SMOTE - Edited Nearest Neighbor)	41
Tabel 4.5 Pseudocode XGBoost (Extreme Gradient Boosting).....	41
Tabel 4.6 Pseudocode LightGBM (Light Gradient Boosting Machine)	42
Tabel 4.7 Pseudocode Perhitungan F1-Score.....	43
Tabel 4.8 Pseudocode Perhitungan G-Mean.....	44
Tabel 4.9 Pseudocode Perhitungan ROC-AUC	44
Tabel 4.10 Distribusi Kelas Data Latih Sebelum dan Setelah Oversampling.....	46
Tabel 4.11 Peringkat Performa Teknik Oversampling pada Dataset Bank Marketing (Rata-Rata XGBoost dan LightGBM)	48
Tabel 4.12 Peringkat Performa Teknik Oversampling pada Dataset Glass Identification (Rata-Rata XGBoost dan LightGBM)	49
Tabel 4.13 Peringkat Performa Teknik Oversampling pada Dataset Musk Version I (Rata-Rata XGBoost dan LightGBM)	50
Tabel 4.14 Hasil Lengkap Kombinasi Teknik Oversampling dan Classifier pada Data Uji (Test Set 20%).....	52
Tabel 4.15 Kombinasi Teknik Oversampling dan Classifier Terbaik per Dataset	54
Tabel 4.16 Kontribusi Teknik Oversampling Terhadap Baseline (Delta F1-Score) per Classifier	56

DAFTAR GAMBAR

Gambar 3.1 Flowchart Teknis Penelitian.....	32
Gambar 3.2 Rancangan Penelitian	33



DAFTAR ISTILAH

<i>Accuracy paradox</i>	Kondisi pada data <i>imbalance</i> di mana model yang memprediksi seluruh sampel sebagai kelas mayoritas tetap menghasilkan akurasi keseluruhan yang tinggi, meskipun gagal total mendeteksi kelas minoritas yang justru paling penting untuk diidentifikasi.
<i>ADASYN (Adaptive Synthetic Sampling)</i>	Teknik <i>oversampling</i> yang mengembangkan SMOTE dengan menghasilkan jumlah sampel sintetis secara adaptif berdasarkan tingkat kesulitan belajar setiap sampel minoritas, di mana sampel yang berada di area lebih sulit (dikelilingi lebih banyak tetangga kelas mayoritas) mendapat prioritas sintesis lebih besar.
<i>Baseline</i>	Kondisi acuan tanpa penerapan teknik <i>oversampling</i> yang disertakan dalam desain eksperimen untuk mengukur kontribusi nyata setiap teknik <i>oversampling</i> terhadap peningkatan performa classifier.
<i>Binary classification</i>	Permasalahan klasifikasi dengan dua kelas target, misalnya nasabah berlangganan deposito atau tidak, yang menjadi jenis tugas klasifikasi pada ketiga dataset dalam penelitian ini.
<i>Borderline</i>	Area pada ruang fitur tempat sampel kelas minoritas dan mayoritas saling tumpang tindih, sehingga sampel di area ini paling rentan terhadap kesalahan klasifikasi dan menjadi target utama sintesis pada Borderline-SMOTE.

<i>Borderline-SMOTE</i>	Varian SMOTE yang hanya menyintesis sampel baru pada sampel minoritas berkategori <i>danger</i> atau <i>borderline</i> , yaitu sampel yang berada di area batas keputusan paling kritis, sehingga upaya penyeimbangan kelas lebih terfokus dibandingkan SMOTE standar.
<i>Boosting</i>	Pendekatan ensemble learning yang membangun model secara bertahap, di mana setiap model baru dilatih untuk memperbaiki kesalahan dari model sebelumnya; XGBoost dan LightGBM merupakan dua algoritma boosting berbasis gradien yang dievaluasi dalam penelitian ini.
<i>Classifier</i>	Algoritma klasifikasi yang digunakan untuk memprediksi kelas target; penelitian ini membatasi evaluasi pada dua classifier boosting modern, yaitu XGBoost dan LightGBM.
<i>Class imbalance</i>	Kondisi distribusi kelas dalam dataset yang tidak proporsional, di mana kelas yang paling penting untuk dideteksi (kelas minoritas) memiliki jumlah sampel jauh lebih sedikit dibandingkan kelas mayoritas.
<i>Convex hull</i>	Wilayah ruang fitur terkecil yang mencakup seluruh titik data minoritas; SMOTE menghasilkan sampel sintetis yang berada di dalam wilayah ini melalui interpolasi linear, bukan sekadar duplikasi data asli.
<i>Cross-validation</i>	Protokol evaluasi model yang membagi data ke dalam beberapa lipatan (<i>fold</i>) untuk melatih dan mengukur performa model secara berulang; penelitian ini menggunakan skema <i>5-fold stratified cross-validation</i> .

<i>Danger (kategori sampel)</i>	Kategori sampel minoritas pada Borderline-SMOTE yang tetangga terdekatnya berasal dari campuran kelas minoritas dan mayoritas, menandakan posisi di area batas keputusan yang kritis.
<i>Exclusive Feature Bundling (EFB)</i>	Inovasi teknis pada LightGBM yang menggabungkan fitur-fitur bersifat <i>mutually exclusive</i> ke dalam satu bundel untuk mengurangi dimensionalitas data secara efektif tanpa kehilangan informasi penting.
<i>F1-Score</i>	Metrik evaluasi berupa rata-rata harmonik antara <i>precision</i> dan <i>recall</i> , digunakan untuk mengukur kemampuan model mendeteksi kelas minoritas karena sensitif terhadap kegagalan pada salah satu dari kedua aspek tersebut.
<i>G-Mean</i>	Metrik evaluasi berupa rata-rata geometris antara <i>sensitivity</i> dan <i>specificity</i> , bernilai mendekati nol apabila model gagal mendeteksi salah satu kelas sama sekali sehingga peka terhadap ketidakseimbangan performa antarkelas.
<i>Gradient-based One-Side Sampling (GOSS)</i>	Inovasi teknis pada LightGBM yang mempertahankan seluruh sampel bergradien besar (lebih informatif) sambil secara acak mereduksi sampel bergradien kecil, sehingga volume data yang diproses berkurang tanpa kehilangan informasi kritikal.
<i>Imbalanced learning</i>	Bidang kajian dalam machine learning yang berfokus pada pengembangan metode untuk menangani permasalahan <i>class imbalance</i> agar model klasifikasi tetap mampu mendeteksi kelas minoritas secara andal.

<i>Imbalance Ratio (IR)</i>	Ukuran rasio jumlah sampel kelas mayoritas terhadap kelas minoritas yang menggambarkan tingkat ketidakseimbangan suatu dataset; penelitian ini menggunakan tiga dataset dengan IR berbeda, yaitu 7,55:1, 15,46:1, dan 1,3:1.
<i>LightGBM (Light Gradient Boosting Machine)</i>	Algoritma <i>gradient boosting</i> yang dirancang untuk efisiensi komputasi tinggi pada dataset besar dan berdimensi tinggi, memperkenalkan mekanisme <i>Gradient-based One-Side Sampling</i> dan <i>Exclusive Feature Bundling</i> , serta menggunakan strategi pertumbuhan pohon <i>leaf-wise</i> .
<i>Noise (sampel)</i>	Sampel data yang tidak merepresentasikan pola sebenarnya dari kelasnya, misalnya sampel minoritas yang seluruh tetangga terdekatnya berasal dari kelas mayoritas; sampel semacam ini berisiko menghasilkan sintesis yang menyesatkan apabila tidak ditangani dengan tepat.
<i>Noise filter</i>	Mekanisme yang membersihkan sampel-sampel ambigu atau bertumpang tindih antarkelas setelah proses oversampling, sebagaimana diterapkan oleh Edited Nearest Neighbor pada SMOTE-ENN.
<i>One-hot encoding</i>	Teknik konversi fitur kategorial menjadi representasi numerik biner, digunakan pada tahap <i>preprocessing</i> sebelum data dimasukkan ke dalam pipeline pelatihan model.
<i>Outlier</i>	Sampel data yang menyimpang signifikan dari pola umum kelasnya; pada teknik <i>oversampling</i> seperti SMOTE dan ADASYN, sintesis dari sampel outlier berisiko memperburuk kualitas batas keputusan classifier.

Oversampling	Pendekatan penanganan <i>class imbalance</i> pada level data yang menambah jumlah sampel kelas minoritas, baik melalui duplikasi maupun pembuatan sampel sintetis, agar distribusi kelas menjadi lebih seimbang sebelum pelatihan model.
Preprocessing	Tahapan pengolahan data sebelum pelatihan model, meliputi penanganan <i>missing values</i> , <i>encoding</i> fitur kategorial, normalisasi fitur numerik, pembagian data latih dan uji, serta penerapan <i>oversampling</i> .
ROC-AUC	Metrik evaluasi yang mengukur kemampuan classifier memisahkan dua kelas pada berbagai ambang batas keputusan (<i>threshold</i>), dihitung berdasarkan kurva <i>Receiver Operating Characteristic</i> .
Safe (kategori sampel)	Kategori sampel minoritas pada Borderline-SMOTE yang sebagian besar tetangga terdekatnya berasal dari kelas minoritas, menandakan posisi yang relatif jauh dari area tumpang tindih kelas dan sudah mudah dipelajari classifier.
Scale_pos_weight	Parameter pada XGBoost dan LightGBM yang memberikan bobot lebih besar pada kesalahan prediksi kelas minoritas selama pelatihan, sebagai mekanisme penanganan <i>class imbalance</i> pada level algoritma.
SMOTE (Synthetic Minority Over-sampling Technique)	Teknik <i>oversampling</i> paling banyak digunakan yang menghasilkan sampel sintetis melalui interpolasi linear antara sampel minoritas dan salah satu tetangga terdekatnya, menjadi fondasi bagi pengembangan berbagai varian <i>oversampling</i> modern.

SMOTE-ENN Teknik *oversampling hybrid* yang menggabungkan SMOTE untuk menghasilkan sampel sintetis dan *Edited Nearest Neighbor* untuk membersihkan sampel yang misklasifikasi oleh tetangga terdekatnya, menghasilkan dataset yang lebih seimbang dan lebih bersih secara spasial.

(Synthetic Minority Over-sampling Technique - Edited Nearest Neighbor)

Stratified cross-validation Skema *cross-validation* yang menjaga proporsi setiap kelas tetap konsisten pada setiap lipatan pelatihan dan validasi, penting untuk memastikan representasi kelas minoritas tidak hilang akibat pembagian data secara acak.

XGBoost (Extreme Gradient Boosting) Algoritma *gradient boosting* yang dioptimasi untuk skalabilitas, efisiensi komputasi, dan akurasi tinggi pada data tabular, dengan regularisasi L1 dan L2 terintegrasi pada fungsi objektifnya serta dukungan penanganan *missing values* secara native.