



**METODE ADAPTIVE RISK-SMOTE DALAM MENANGANI
KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT
PERANGKAT LUNAK**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Sarjana Strata-1 Ilmu Komputer**

Oleh

AZKA NABIL TAQI

NIM 2211016210007

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

JUNI 2026



**METODE ADAPTIVE RISK-SMOTE DALAM MENANGANI
KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT
PERANGKAT LUNAK**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Sarjana Strata-1 Ilmu Komputer**

Oleh

AZKA NABIL TAQI

NIM 2211016210007

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

JUNI 2026

SKRIPSI

METODE ADAPTIVE RISK-SMOTE DALAM MENANGANI KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT PERANGKAT LUNAK

Oleh

AZKA NABIL TAQI

NIM 2211016210007

Telah dipertahankan di depan Dosen Penguji pada tanggal 28 April 2026.

Susunan Dosen Penguji:

Pembimbing Utama



Radityo Adi Nugroho, S.T., M.Kom.
NIP. 198212042008011006

Penguji

1.



Setyo Wahyu Saputro, S.Kom., M.Kom.
NIP. 198808072023211027

Pembimbing Pendamping



Muhammad Utqan Mazdadi, S.Kom,
M.Kom.
NIP. 199006122019031013

2.


Andi Farmadi, S.Si, M.T.
NIP. 197307252008011006

Banjarbaru, 9 Juni 2026

Koordinator Program Studi Ilmu Komputer



Dr. Kartini, S.Kom., M.Kom.
NIP. 198704212012122003

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.

Banjarbaru, 9 Juni 2026



Azka Nabil Taqi
NIM. 2211016210007

ABSTRAK

METODE ADAPTIVE RISK-SMOTE DALAM MENANGANI KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT PERANGKAT LUNAK (Oleh: Azka Nabil Taqi; Pembimbing: Radityo Adi Nugroho, S.T, M.Kom dan Muhammad Itqan Mazdadi, S.Kom, M.Kom; 2026; 65 halaman)

Prediksi cacat perangkat lunak menghadapi tantangan utama berupa ketidakseimbangan kelas, di mana model klasifikasi cenderung mengabaikan kelas minoritas. Metode *oversampling* seperti SMOTE menghasilkan sampel sintetis secara seragam tanpa mempertimbangkan karakteristik risiko masing-masing sampel, sehingga berpotensi menghasilkan sampel yang tidak informatif. Penelitian ini mengusulkan *Adaptive Risk-SMOTE* (AR-SMOTE), metode *oversampling* baru yang mengombinasikan *entropi Shannon* dan kepadatan mayoritas lokal sebagai skor risiko untuk mengarahkan sintesis sampel secara selektif pada kategori minoritas yang paling informatif. Pengujian dilakukan pada dataset AEEEM (EQ, JDT, LC, ML, PDE) menggunakan enam algoritma klasifikasi dengan skema *5-fold stratified cross-validation* dan metrik AUC.

Hasil menunjukkan AR-SMOTE unggul pada Logistic Regression dan *Neural Network* di 3 dari 5 dataset, dengan peningkatan terbesar pada LC untuk Logistic Regression (+0,0397 terhadap SMOTE) dan JDT untuk *Neural Network* (+0,0239 terhadap SMOTE). Secara rata-rata keseluruhan, AR-SMOTE (0,7768) mengungguli SMOTE standar (0,7753), Borderline-SMOTE (0,7748), dan ADASYN (0,7748). Namun AR-SMOTE kurang optimal pada KNN dan SVM. AR-SMOTE terbukti menjadi alternatif *oversampling* yang selektif dan adaptif, dengan efektivitas tertinggi pada model yang bergantung pada kualitas distribusi data latih.

Kata Kunci: *Software Defect Prediction*, Ketidakseimbangan Kelas, AR-SMOTE, *Oversampling*, *Entropi Shannon*.

ABSTRACT

ADAPTIVE RISK-SMOTE METHOD IN HANDLING DATA IMBALANCE FOR SOFTWARE DEFECT PREDICTION (By: Azka Nabil Taqi; Supervisor: Radityo Adi Nugroho, S.T, M.Kom and Muhammad Itqan Mazdadi, S.Kom, M.Kom; 2026; 65 pages)

Software defect prediction faces a critical challenge in class imbalance, where classifiers tend to neglect the minority class. Oversampling methods such as SMOTE generate synthetic samples uniformly without considering individual sample risk, potentially producing uninformative samples. This study proposes Adaptive Risk-SMOTE (AR-SMOTE), a novel oversampling method that combines Shannon entropy and local majority density as a risk score to selectively direct synthetic sample generation toward the most informative minority category. Experiments were conducted on the AEEEM dataset (EQ, JDT, LC, ML, PDE) using six classifiers under a 5-fold stratified cross-validation scheme with AUC as the evaluation metric.

Results show that AR-SMOTE achieves the highest AUC for both Logistic Regression and Neural Network in 3 out of 5 datasets, with the largest improvements on LC for Logistic Regression (+0.0397 over SMOTE) and JDT for Neural Network (+0.0239 over SMOTE). In overall average AUC, AR-SMOTE (0.7768) outperforms standard SMOTE (0.7753), Borderline-SMOTE (0.7748), and ADASYN (0.7748), though it performs less favorably for KNN and SVM. Overall, AR-SMOTE proves to be a selective and adaptive oversampling alternative with its greatest effectiveness on models that rely on training data distribution quality.

Keywords: *Software Defect Prediction, Class Imbalance, AR-SMOTE, Oversampling, Shannon Entropy.*

PRAKATA

Alhamdulillah rabbil ‘alamin, puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi berjudul **“METODE ADAPTIVE RISK-SMOTE DALAM MENANGANI KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT PERANGKAT LUNAK”** sebagai syarat menyelesaikan pendidikan S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

Penulis menyampaikan terima kasih kepada:

1. Bapak Radityo Adi Nugroho, S.T., M.Kom., selaku pembimbing utama, atas bimbingan dan arahannya.
2. Bapak Muhammad Itqan Mazdadi, S.Kom, M.Kom., selaku pembimbing pendamping, atas saran konstruktifnya.
3. Bapak Setyo Wahyu Saputro, S.Kom., M.Kom. dan Bapak Andi Farmadi, S.Si, M.T., selaku penguji, atas kritik dan sarannya.
4. Ibu Dwi Kartini, S.Kom., M.Kom., selaku Ketua Program Studi.
5. Seluruh dosen dan staf Prodi Ilmu Komputer FMIPA ULM.
6. Semua pihak lain yang turut membantu.

Penulis menyadari adanya kekurangan dalam penyusunan skripsi ini dan terbuka terhadap kritik dan saran demi perbaikan di masa mendatang. Semoga karya ini bermanfaat dan mendapat ridha dari Allah Subhanahu wa Ta'ala.

Banjarbaru, 9 Juni 2026



Azka Nabil Taqi

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
PERNYATAAN.....	iii
ABSTRAK.....	iv
<i>ABSTRACT</i>	v
PRAKATA.....	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xii
BAB I PENDAHULUAN	
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
BAB II TINJAUAN PUSTAKA	
2.1 Kajian Terdahulu.....	6
2.2 <i>Software Defect Prediction (SDP)</i>	8
2.3 <i>Class Imbalance dan Oversampling</i>	9
2.4 Algoritma Klasifikasi	10
2.4.1 <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	10
2.4.2 <i>Borderline SMOTE</i>	12
2.4.3 <i>Adaptive Syntethic Sampling (ADASYN)</i>	14
2.4.4 <i>SVM SMOTE</i>	16
2.4.5 <i>Adaptive Weighted SMOTE (AWSMOTE)</i>	19
2.5 <i>Area Under the ROC Curve (AUC)</i>	22
2.6 Potensi Metode <i>Oversampling</i> Berbasis Bobot.....	23

BAB III METODE PENELITIAN

3.1	Prosedur Penelitian.....	25
3.2	Alat Penelitian.....	25
3.3	Bahan Penelitian.....	26
3.4	Alur Penelitian	26
3.4.1	Perhitungan Skor Risiko (<i>Risk Score Computation</i>).....	28
3.4.2	Evaluasi dan Seleksi Kategori.....	30
3.4.3	<i>Oversampling</i> Adaptif Berbasis Risiko.....	31
3.4.4	Konfigurasi Parameter Implementasi.....	31
3.4.5	Pseudocode Algoritma AR-SMOTE.....	33
3.4.6	Analisis Kompleksitas Waktu (Big-O)	34

BAB IV HASIL DAN PEMBAHASAN

4.1	Data Understanding.....	38
4.1.1	Statistik Deskriptif Dataset	38
4.1.2	Analisis Distribusi Kelas dan Tingkat Ketidakseimbangan.....	39
4.1.3	Analisis Distribusi Fitur dan Korelasi.....	40
4.1.4	Temuan Kunci <i>Data Understanding</i>	40
4.2	<i>Data Preparation</i>	41
4.2.1	Hasil <i>Encoding Label</i> dan Verifikasi Integritas	41
4.2.2	Dampak Normalisasi terhadap Distribusi Data.....	42
4.2.3	Hasil Partisi <i>Stratified K-Fold</i>	42
4.3	<i>Modeling</i>	43
4.3.1	Distribusi Skor Risiko Sampel Minoritas	43
4.3.2	Seleksi Kategori Terbaik per Dataset.....	45
4.3.3	Seleksi Kategori Terbaik per Dataset.....	47
4.4	Evaluasi	49
4.4.1	Hasil Evaluasi <i>Gradient Boosting</i>	49
4.4.2	Hasil Evaluasi <i>Random Forest</i>	49
4.4.3	Hasil Evaluasi KNN	50

4.4.4 Hasil Evaluasi <i>Logistic Regression</i>	51
4.4.5 Hasil Evaluasi MLP dan SVM	52
4.4.6 Perbandingan Komperhensif Lintas Metode dan Dataset	53
4.4.7 Uji Signifikansi Statistik	55
4.5 Pembahasan	57
BAB V PENUTUP	
5.1. Kesimpulan	60
5.2. Saran	61
DAFTAR PUSTAKA	63
LAMPIRAN	

DAFTAR TABEL

Tabel 1 Parameter Global AR-SMOTE	32
Tabel 2 Konfigurasi <i>Library scikit-learn</i> dan <i>imbalanced-learn</i> per Tahapan	32
Tabel 3 Analisis Kompleksitas Waktu per Tahapan AR-SMOTE.....	35
Tabel 4 Perbandingan Kompleksitas Waktu Metode <i>Oversampling</i>	36
Tabel 5 Statistik Deskriptif Agregat Dataset AEEEM.....	38
Tabel 6 Analisis Ketidakseimbangan Kelas Dataset AEEEM	39
Tabel 7 Hasil Representatif <i>label encoding</i> pada Dataset.....	41
Tabel 8 Perbandingan Statistik Sebelum dan Sesudah Normalisasi (Sampel Representatif)	42
Tabel 9 Distribusi Sampel per <i>Fold</i> pada Dataset LC (<i>5-Fold Stratified CV</i>)	43
Tabel 10 Statistik Distribusi Skor Risiko Sampel (Fold 1).....	44
Tabel 11 Detail Distribusi Sampel per Kategori AR-SMOTE (<i>Fold 1</i>)	45
Tabel 12 Skor <i>Regularized AUC</i> per Kategori dan K^* Terpilih (<i>Fold 1</i>).....	45
Tabel 13 Kategori Risiko Terpilih (K^*) AR-SMOTE per Dataset	46
Tabel 14 Hasil <i>Oversampling</i> Adaptif AR-SMOTE per Dataset.....	48
Tabel 15 Nilai AUC pada <i>Gradient Boosting</i>	49
Tabel 16 Nilai AUC pada <i>Random Forest</i>	50
Tabel 17 Nilai AUC pada KNN	50
Tabel 18 Nilai AUC pada <i>Logistic Regression</i>	51
Tabel 19 Nilai AUC pada <i>Neural Network</i> (MLP)	52

Tabel 20 Nilai AUC pada SVM	53
Tabel 21 Perbandingan <i>Classifier</i> AR-SMOTE	53
Tabel 22 Rata-rata AUC Seluruh Dataset per Metode dan Algoritma	54
Tabel 23 Hasil Uji Wilcoxon Signed-Rank AR-SMOTE vs Baseline ($n = 30$, $\alpha = 0,05$)	56

DAFTAR GAMBAR

Gambar 1 Konsep <i>Workflow</i> Algoritma AR-SMOTE	27
Gambar 2 Grafik Nilai AUC pada <i>Logistic Regression</i>	51
Gambar 3 Grafik Nilai AUC pada <i>Neural Network</i>	52
Gambar 4 Grafik Perbandingan Nilai AUC AR-SMOTE dan Metode <i>Baseline</i> ..	55