



**PERBANDINGAN *SMOTE-VARIANTS* UNTUK MENGATASI
KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT *SOFTWARE***

Skripsi

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

Oleh

MUIZA RAHMAN

NIM. 1911016310018

**PROGRAM STUDI S1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

NOVEMBER 2023



**PERBANDINGAN *SMOTE-VARIANTS* UNTUK MENGATASI
KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT *SOFTWARE***

Skripsi

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

Oleh

MUIZA RAHMAN

NIM. 1911016310018

**PROGRAM STUDI S1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
NOVEMBER 2023**

SKRIPSI

**PERBANDINGAN *SMOTE-VARIANTS* UNTUK MENGATASI
KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT *SOFTWARE***

Oleh :

MUIZA RAHMAN

1911016310018

Telah dipertahankan di depan Dosen Penguji pada tanggal 27 September 2023

Susunan Dosen Penguji :

Pembimbing I



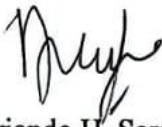
Mohammad Reza Faisal, S.T., M.T., Ph.D.
NIP. 197612202008121001

Dosen Penguji I



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

Pembimbing II



Triando H. Saragih, S.Kom., M.Kom.
NIP. 199308242019031012

Dosen Penguji II



Friska Abadi, S.Kom., M.Kom.
NIP. 198809132023211010



Banjarnbaru, 15 November 2023
Ketua Program Studi Ilmu Komputer

Irwan Budiman S.T., M.Kom
NIP. 197703252008121001

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, 15 November 2023

Yang menyatakan,



Muiza Rahman

NIM. 1911016310018

ABSTRAK

PERBANDINGAN *SMOTE-VARIANTS* UNTUK MENGATASI KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT *SOFTWARE*

(Oleh : Muiza Rahman; Pembimbing: Mohammad Reza Faisal, S.T., M.T. Ph.D dan Triando Hamonangan Saragih, S.Kom., M.Kom; 2023; 46 halaman)

Ketidakseimbangan data merupakan salah satu masalah penelitian prediksi cacat *software*. Ketidakseimbangan data dapat membuat model bias kepada kelas mayoritas, sementara kebanyakan dataset untuk prediksi cacat *software*, kelas cacat yang ingin diprediksi merupakan kelas minoritas. Metode umum yang digunakan untuk mengatasi ini adalah pendekatan *level* data seperti *SMOTE-Variants* ataupun pendekatan *level* algoritma seperti *Bagging*. Penelitian ini menggunakan pendekatan *level* data *SMOTE-Variants* yaitu *SMOTE*, *Borderline-SMOTE*, *SMOTE-ENN* dan *SMOTE-Tomek-Links*. Sementara, pendekatan *level* algoritma yang digunakan adalah *Bagging (Decision Tree)* atau *Bagging* dengan *base classifier Decision Tree*. Hasilnya model dengan *SMOTE-Tomek-Links* dan *Bagging (Decision Tree)* mendapatkan skor rata-rata *AUC* terbaik sebesar 0,8009. Sementara, *SMOTE-ENN* dan *Bagging (Decision Tree)* mendapatkan skor rata-rata *recall* terbaik sebesar 0,6480.

Kata kunci: Ketidakseimbangan data, Prediksi cacat *software*, *Bagging*, *SMOTE-Variants*

ABSTRACT

COMPARISON OF SMOTE-VARIANTS TO OVERCOME DATA IMBALANCE IN SOFTWARE DEFECT PREDICTION

(by : Muiza Rahman; *Supervisor:* Mohammad Reza Faisal, S.T., M.T. Ph.D and Triando Hamonangan Saragih, S.Kom., M.Kom; 2023; 46 pages)

Data imbalance is one of the problems of software defect prediction research. Data imbalance can make the model biased towards the majority class, while in most of the datasets for software defect prediction, the defect class to be predicted is the minority class. Common methods used to address this are data-level approach such as SMOTE-Variants or classification-level approach such as Bagging. This research uses the SMOTE-Variants data-level approach, namely SMOTE, Borderline-SMOTE, SMOTE-ENN and SMOTE-Tomek-Links. Meanwhile, the algorithm level approach used is Bagging (Decision Tree) or Bagging with Decision Tree base classifier. As a result, the model with SMOTE-Tomek-Links and Bagging (Decision Tree) gets the best average AUC score of 0.8009. Meanwhile, SMOTE-ENN and Bagging (Decision Tree) get the best average recall score of 0.6480.

Keywords: *Data imbalance, Software Defect Prediction, Bagging, SMOTE-Variants*

PRAKATA

Puji syukur penulis panjatkan ke Tuhan kita Yang Maha Esa karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul **PERBANDINGAN *SMOTE-VARIANTS* UNTUK MENGATASI KETIDAKSEIMBANGAN DATA PADA PREDIKSI CACAT *SOFTWARE*** untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

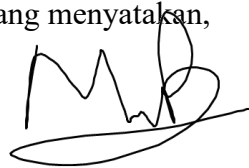
Pada lembar ini penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut:

1. Keluarga yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini.
2. Bapak Mohammad Reza Faisal, S.T., M.T. Ph.D selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Bapak Triando Hamonangan Saragih, S.Kom., M.Kom selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Irwan Budiman S.T., M.Kom selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
6. Teman-teman dan sahabat-sahabat keluarga Ilmu Komputer angkatan 2019 yang memberikan dukungan dan selalu mengingatkan serta mendoakan dalam proses mengerjakan skripsi.
7. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna, namun penulis mengharapkan bantuan berupa saran dan kritik yang membangun dari semua pihak demi kesempurnaan dan mutu penulisan skripsi ini. Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, 15 November 2023

Yang menyatakan,



Muiza Rahman
NIM. 1911016310018

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN	ii
PERNYATAAN	iii
ABSTRAK	iv
ABSTRACT	v
PRAKATA	vi
DAFTAR ISI	viii
DAFTAR TABEL	x
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiii
BAB I	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
BAB II	4
2.1 Kajian Terdahulu	4
2.2 Prediksi cacat Software	7
2.3 Ketidakseimbangan data	8
2.4 Dataset AEEEM	8
2.5 <i>SMOTE</i>	9
2.6 <i>Borderline-SMOTE</i>	11
2.7 <i>SMOTE-ENN</i>	13
2.8 <i>SMOTE-Tomek-Links</i>	14
2.9 <i>Decision Tree</i>	15
2.10 <i>Bagging</i>	16
2.11 <i>Cross Validation</i>	17

2.12 Evaluasi	18
BAB III	21
3.1 Alat Penelitian	21
3.2 Bahan Penelitian	21
3.3 Variabel Penelitian	21
3.4 Prosedur Penelitian	22
BAB IV	24
4.1 Hasil	24
4.1.1 Pengumpulan Data	24
4.1.2 <i>Encoding</i> Label	25
4.1.3 Pembagian Data	25
4.1.4 Penyeimbangan Data	28
4.1.5 Evaluasi	31
4.2 Pembahasan	33
BAB V	40
5.1 Kesimpulan	40
5.2 Saran	40
DAFTAR PUSTAKA	41
LAMPIRAN	48

DAFTAR TABEL

Tabel 1 . Keaslian Penelitian	6
Tabel 2 . Perancangan Penelitian	7
Tabel 3 . Dataset <i>AEEEM</i>	9
Tabel 4 . Contoh salah satu dataset.	9
Tabel 5 . Sampel Acak data <i>SMOTE</i>	10
Tabel 6 . Data sintetis <i>SMOTE</i>	11
Tabel 7 . Sampel acak <i>Borderline-SMOTE</i>	12
Tabel 8 . Tetangga terdekat sampel acak <i>Borderline-SMOTE</i>	12
Tabel 9 . Data sintetis <i>Borderline-SMOTE</i>	13
Tabel 10 . Sampel acak <i>SMOTE-ENN</i>	14
Tabel 11 . Sampel acak <i>SMOTE-Tomek-Links</i>	15
Tabel 12 . Hasil contoh <i>Bagging</i>	17
Tabel 13 . <i>Confusion Matrix</i>	19
Tabel 14 . Performa <i>AUC</i>	20
Tabel 15 . Jumlah data pada dataset <i>AEEEM</i>	24
Tabel 16 . Encoding dataset <i>AEEEM</i>	25
Tabel 17 . Pembagian Dataset <i>EQ</i>	25
Tabel 18 . Pembagian Dataset <i>JDT</i>	26
Tabel 19 . Pembagian Dataset <i>LC</i>	26
Tabel 20 . Pembagian Dataset <i>ML</i>	27
Tabel 21 . Pembagian Dataset <i>PDE</i>	27
Tabel 22 . Penyeimbangan Dataset <i>EQ</i> dengan <i>SMOTE-Variants</i>	28
Tabel 23 . Penyeimbangan Dataset <i>JDT</i> dengan <i>SMOTE-Variants</i>	29
Tabel 24 . Penyeimbangan Dataset <i>LC</i> dengan <i>SMOTE-Variants</i>	29
Tabel 25 . Penyeimbangan Dataset <i>ML</i> dengan <i>SMOTE-Variants</i>	30
Tabel 26 . Penyeimbangan Dataset <i>PDE</i> dengan <i>SMOTE-Variants</i>	30
Tabel 27 . Hasil Evaluasi dataset <i>EQ</i> menggunakan <i>recall</i> , <i>FPR</i> dan <i>AUC</i>	31
Tabel 28 . Hasil Evaluasi dataset <i>JDT</i> menggunakan <i>recall</i> , <i>FPR</i> dan <i>AUC</i>	31
Tabel 29 . Hasil Evaluasi dataset <i>LC</i> menggunakan <i>recall</i> , <i>FPR</i> dan <i>AUC</i>	32

Tabel 30 . Hasil Evaluasi dataset *ML* menggunakan *recall*, *FPR* dan *AUC*. 32

Tabel 31 . Hasil Evaluasi dataset *PDE* menggunakan *recall*, *FPR* dan *AUC*.32

DAFTAR GAMBAR

Gambar 1 . Ilustrasi <i>SMOTE</i> (Pei et al., 2022).....	10
Gambar 2 . <i>Borderline-SMOTE</i> (Pei et al., 2022).....	12
Gambar 3 . Ilustrasi <i>SMOTE-ENN</i> (Lin & Zeng, 2021).....	13
Gambar 4 . Ilustrasi <i>SMOTE-Tomek-Links</i> (Cho et al., 2022).....	15
Gambar 5 . Ilustrasi <i>Decision Tree</i>	16
Gambar 6 . Ilustrasi <i>Bagging</i> (Yang et al., 2019).....	17
Gambar 7 . Ilustrasi <i>K-fold Cross Validation</i> (Ren et al., 2019).....	18
Gambar 8 . Alur Prosedur Penelitian.....	22
Gambar 9 . Perbandingan Skor <i>Recall</i>	35
Gambar 10 . Perbandingan Skor <i>FPR</i>	35
Gambar 11 . Perbandingan Skor Rata-rata <i>Recall</i> dan <i>FPR</i>	36
Gambar 12 . Perbandingan Skor <i>AUC</i>	38
Gambar 13 . Perbandingan Skor rata-rata <i>AUC</i>	39

DAFTAR LAMPIRAN

Lampiran 1 Perhitungan <i>Recall</i> dan <i>FPR</i> (<i>Bagging (Decision Tree)</i>)	49
Lampiran 2 Perhitungan <i>Recall</i> dan <i>FPR</i> (<i>SMOTE + Bagging (Decision Tree)</i>)	53
Lampiran 3 Perhitungan <i>Recall</i> dan <i>FPR</i> (<i>Borderline- SMOTE + Bagging (Decision Tree)</i>)	57
Lampiran 4 Perhitungan <i>Recall</i> dan <i>FPR</i> (<i>SMOTE-ENN + Bagging (Decision Tree)</i>)	61
Lampiran 5 Perhitungan <i>Recall</i> dan <i>FPR</i> (<i>SMOTE-Tomek-Links + Bagging (Decision Tree)</i>)	65
Lampiran 6 Source Code	69
Lampiran 7 Riwayat Hidup	74