



SKRIPSI

ANALISIS PERBANDINGAN PERFORMANSI TEKNIK RESAMPLING DATA SMOTE DAN SMOTEN PADA KLASIFIKASI MULTICLASS CURAH HUJAN HARIAN MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST)

NAJWA LAILA ANGGRAINI

NPM 21081010191

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S.Si., M.T

Achmad Junaidi, S.Kom., M.Kom

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2025



SKRIPSI

**ANALISIS PERBANDINGAN PERFORMANSI TEKNIK
RESAMPLING DATA SMOTE DAN SMOTEN PADA
KLASIFIKASI MULTICLASS CURAH HUJAN HARIAN
MENGUNAKAN EXTREME GRADIENT BOOSTING
(XGBOOST)**

NAJWA LAILA ANGGRAINI

NPM 21081010191

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S.Si., M.T.

Achmad Junaidi, S.Kom., M.Kom.

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2025

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

ANALISIS PERBANDINGAN PERFORMANSI TEKNIK RESAMPLING DATA SMOTE DAN SMOTEN PADA KLASIFIKASI MULTICLASS CURAH HUJAN HARIAN MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST)

Oleh:

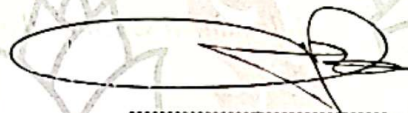
NAJWA LAILA ANGGRAINI

NPM. 21081010191

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Informatika
Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada
Tanggal 16 Mei 2025.

Dr. Basuki Rahmat, S.Si., M.T.

NIP. 19690723 2021211 002



(Pembimbing I)

Achmad Junaidi, S.Kom., M.Kom.

NPT. 3 7811 04 0199 1



(Pembimbing II)

Fawwaz Ali Akbar, S.Kom., M.Kom.

NIP. 19920317 201803 1 002



(Ketua Penguji)

Firza Prima Aditiawan, S.Kom., M.T.I.

M.C.F, M.O.S.

NIP. 19860523 202121 1 003



(Anggota Penguji)

Mengetahui,
Dekan Fakultas Ilmu Komputer



Prof. Dr. Ir. Novirina Hendrasarie, MT.

NIP. 19681126 199403 2 001

Halaman ini sengaja dikosongkan

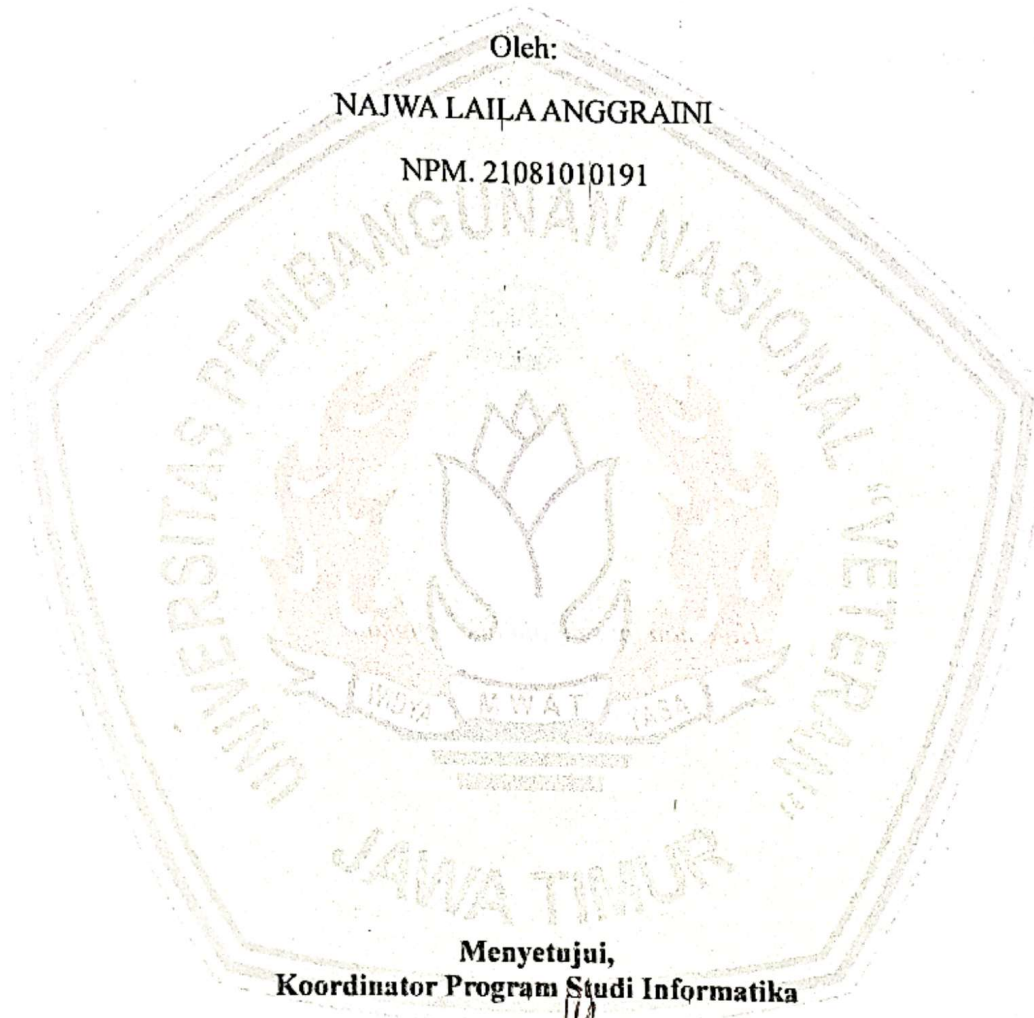
LEMBAR PERSETUJUAN

**ANALISIS PERBANDINGAN PERFORMANSI TEKNIK RESAMPLING DATA
SMOTE DAN SMOTEN PADA KLASIFIKASI MULTICLASS CURAH HUJAN
HARIAN MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST)**

Oleh:

NAJWA LAILA ANGGRAINI

NPM. 21081010191



**Menyetujui,
Koordinator Program Studi Informatika**

Fetty Tri Anggraeny, S.Kom., M.Kom.
NIP. 19820211 2021212 005

Halaman ini sengaja dikosongkan

PERNYATAAN BEBAS PLAGIASI

Yang bertandatangan di bawah ini:

Nama Mahasiswa : Najwa Laila Anggraini
NPM : 21081010191
Program Studi : Informatika
Fakultas : Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila adikemudian hari ditemukan indikasi plagiat pada Skripsi ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.



Surabaya, 16 Mei 2025

Yang Membuat Pernyataan,



NAJWA LAILA ANGGRAINI

NPM. 21081010191

Halaman ini sengaja dikosongkan

ABSTRAK

Nama Mahasiswa/NPM : Najwa Laila Anggraini/21081010191

Judul : Analisis Perbandingan Performansi Teknik Resampling Data SMOTE dan SMOTEN Pada Klasifikasi Multiclass Curah Hujan Harian Menggunakan Extreme Gradient Boosting (XGBoost)

Dosen Pembimbing : 1. Dr. Basuki Rahmat, S.Si., M.T.
2. Achmad Junaidi, S.Kom., M.Kom.

Dalam membangun sistem klasifikasi yang akurat diperlukan penanganan dalam banyak faktor, salah satunya masalah ketidakseimbangan data. Beberapa teknik *resampling* data diciptakan untuk memenuhi kebutuhan dalam berbagai jenis dataset. Penelitian ini berfokus pada performansi teknik *resampling* untuk meningkatkan keakuratan model XGBoost dalam melakukan klasifikasi melalui pendekatan SMOTE dan SMOTEN.

Penelitian ini menggunakan dataset sebanyak 5.804 data harian curah hujan pada Provinsi Jawa Timur dengan 8 kolom bertipe numerik dan 1 kolom bertipe kategorikal. Dataset kemudian dilakukan *training* menggunakan *cross validation* dengan 5 dan 10 *fold*, serta penggunaan nilai *learning rate* yang berbeda. Tujuannya untuk melihat ketahanan dan stabilitas dari kedua teknik *resampling* yang digunakan. Hasil yang diperoleh menunjukkan bahwa baik SMOTE maupun SMOTEN dapat meningkatkan nilai akurasi dari model *baseline* tanpa *resampling*.

Dalam pengujiannya, seluruh eksperimen dari SMOTEN memiliki nilai akurasi paling tinggi dan paling stabil yaitu 92,92% membutuhkan 37,2 detik waktu komputasi, kemudian hasil akurasi dari SMOTE 90,58%, dengan waktu komputasi lebih lama 8,4 detik dari SMOTEN yaitu 45,6 detik. Kedua skenario ini terbukti dapat meningkatkan 15,22% hingga 17,56% akurasi dalam penggunaan dataset asli.

Kata Kunci: Curah Hujan, Klasifikasi, XGBoost, SMOTE, SMOTEN, Pembelajaran Mesin.

Halaman ini sengaja dikosongkan

ABSTRACT

Nama Mahasiswa/NPM : Najwa Laila Anggraini/21081010191

Judul : Analisis Perbandingan Performansi Teknik Resampling Data SMOTE dan SMOTEN Pada Klasifikasi Multiclass Curah Hujan Harian Menggunakan Extreme Gradient Boosting (XGBoost)

Dosen Pembimbing : 1. Dr. Basuki Rahmat, S.Si., M.T.
2. Achmad Junaidi, S.Kom., M.Kom.

The development of an accurate classification system necessitates the consideration of numerous factors, including the issue of data imbalance. A range of data resampling techniques have been developed to address the diverse requirements of different data sets. The present study examines the efficacy of resampling techniques in enhancing the precision of the XGBoost model's classification capabilities, employing the SMOTE and SMOTEN methodologies.

The present study utilizes a dataset comprising 5,804 daily rainfall observations in East Java Province, encompassing eight numerical and one categorical column. Subsequently, training is executed employing cross validation with 5 and 10 folds, alongside the utilization of disparate learning rate values. The objective is to assess the robustness and stability of the two resampling techniques employed. The findings indicate that the SMOTE and SMOTEN algorithms are capable of enhancing the accuracy of a baseline model without the necessity of resampling.

In the experimental framework, all SMOTEN-based experiments exhibited the highest and most stable accuracy value of 92.92%, requiring 37.2 seconds of computation time. Conversely, the accuracy result from SMOTE demonstrated 90.58%, requiring a longer computation time of 8.4 seconds (45.6 seconds). It has been demonstrated that both scenarios are effective in enhancing the accuracy of the original dataset, with improvements ranging from 15.22% to 17.56%.

Keywords: rainfall, classification, XGBoost, SMOTE, SMOTEN, and learning.

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Syukur Alhamdulillah, hanya kepada Tuhan Yang Maha Esa, sehingga penulis dapat menyelesaikan skripsi ini sebagai salah satu syarat kelulusan di Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jawa Timur. Semua tidak lepas pihak – pihak yang memberi dukungan selama pengerjaan skripsi. Untuk itu penulis mengucapkan terimakasih kepada:

1. Bapak Basuki Rahmat, S.Si., M.T. selaku Dosen Pembimbing I dan Bapak Achmad Junaidi, S.Kom., M.Kom. selaku Dosen Pembimbing II yang telah meluangkan waktu dan tenaganya memberikan bimbingan dan motivasi kepada penulis.
2. Dosen – dosen Program Studi Informatika yang memberikan ilmu semasa perkuliahan.
3. Ayah dan Ibu tercinta yang senantiasa mengusahakan dan percaya kepada penulis serta selalu memberikan dukungan dalam kondisi apapun sehingga penulis dapat menyelesaikan masa studi tepat waktu.
4. Adik tersayang, atas semangat dan dukungan yang diberikan.
5. NPM 21081010231, selaku partner terbaik yang selalu menemani dan meyakinkan penulis selama pengerjaan skripsi.
6. Teman – teman yang tidak bisa penulis sebutkan satu persatu yang telah menemani penulis dari awal perkuliahan hingga akhir.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk perbaikan dimasa yang akan datang.

Surabaya, 16 Mei 2025
Penulis,


Najwa Laila Anggraini

Halaman ini sengaja dikosongkan

DAFTAR ISI

LEMBAR PENGESAHAN.....	ii
LEMBAR PERSETUJUAN	iv
PERNYATAAN BEBAS PLAGIASI	vi
ABSTRAK.....	viii
ABSTRACT	x
KATA PENGANTAR	xii
DAFTAR ISI	xiv
DAFTAR GAMBAR	xvi
DAFTAR TABEL	xvii
DAFTAR KODE.....	xviii
BAB I PENDAHULUAN	1
1.1 LATAR BELAKANG.....	1
1.2 RUMUSAN MASALAH	4
1.3 TUJUAN PENELITIAN	4
1.3.1. Tujuan Umum.....	5
1.3.2. Tujuan Khusus.....	5
1.4 MANFAAT PENELITIAN	5
1.5 BATASAN MASALAH	6
BAB II TINJAUAN PUSTAKA	7
2.1 PENELITIAN TERDAHULU.....	7
2.2 LANDASAN TEORI	12
2.2.1. Extreme Gradient Boosting (XGBoost)	12
2.2.2. <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	15
2.2.3. <i>Synthetic Minority Oversampling for Nominal (SMOTEN)</i>	16
2.2.4. <i>Imbalanced Dataset</i>	17
2.2.5. Confusion Matriks.....	18
2.2.6. Curah Hujan	19
2.2.7. Tekanan Udara.....	20
2.2.8. Kelembapan Udara	20
BAB III DESAIN DAN IMPLEMENTASI SISTEM.....	22
3.1 METODE PENELITIAN	22
3.2 IDENTIFIKASI MASALAH.....	22
3.3 STUDI LITERATUR	22
3.4 PENGUMPULAN DATA.....	23

3.5	DESAIN SISTEM	24
3.6	PREPROCESSING DATA	25
3.6.1.	<i>Data Cleaning</i>	25
3.6.2.	Standarisasi	28
3.6.3.	<i>Synthetic Minority Oversampling Technique (SMOTE)</i>	31
3.6.4.	<i>Synthetic Minority Oversampling For Nominal (SMOTEN)</i>	34
3.7	XGBOOST MODEL	37
3.8	EVALUASI HASIL	42
3.9	SKENARIO PENGUJIAN	44
BAB IV	PENGUJIAN DAN ANALISA	48
4.1.	PREPROCESSING	48
4.1.1.	Modul	48
4.1.2.	Load Data	49
4.1.3.	Data Cleaning	51
4.1.4.	Standarisasi	53
4.2.	SKENARIO PENGUJIAN	54
4.2.1.	Skenario Pengujian 1 (Baseline Model)	54
4.2.1.1.	Mendefinisikan Model Xgboost	55
4.2.1.2.	Cross Validation Training & Evaluation	56
4.2.2.	Skenario Pengujian 2 (SMOTE Dataset)	57
4.2.2.1.	Teknik Resampling SMOTE	58
4.2.3.	Skenario Pengujian 3 (SMOTEN Dataset)	59
4.2.3.1.	Teknik Resampling SMOTEN	60
4.3.	HASIL PENGUJIAN	61
4.3.1.	Hasil Pengujian Skenario 1	61
4.3.2.	Hasil Pengujian Skenario 2 (SMOTE Dataset)	64
4.3.3.	Hasil Pengujian Skenario 3 (SMOTEN Dataset)	65
4.4.	PERBANDINGAN	69
4.4.1.	Sebaran Dataset	69
4.4.2.	Akurasi	70
4.4.3.	Waktu Komputasi	76
BAB V	PENUTUP	80
5.1.	KESIMPULAN	80
5.2.	SARAN	81
DAFTAR PUSTAKA		83

DAFTAR GAMBAR

Gambar 2.1. Flowchart XGBoost	13
Gambar 3.1. Kerangka Pemikiran.....	22
Gambar 3.2. Desain Sistem.....	24
Gambar 3.3. Tahapan Preprocessing.....	25
Gambar 3.4. Tahapan Handling Missing Value	26
Gambar 3.5. Tahapan Standarisasi data menggunakan Z-Score.....	28
Gambar 3.6. Menampilkan informasi dari dataset menggunakan library python.....	31
Gambar 3.7. Diagram alur implementasi SMOTE dan SMOTEN	32
Gambar 3.8. Gradient Boosting proses training	37
Gambar 3.9. Diagram alur menghitung matriks evaluasi	42
Gambar 3.10. Eksperimen pengujian.....	45
Gambar 4.1. Website pengumpulan data	50
Gambar 4.2. Isi data harian	50
Gambar 4.3. Penggabungan dataset.....	50
Gambar 4.4. Informasi tipe data	51
Gambar 4.5. Jumlah nilai null.....	52
Gambar 4.6. Tampilan dataset setelah penyesuaian fitur.....	52
Gambar 4.7. Hasil setelah pengisian nilai hilang.....	53
Gambar 4.8. Hasil standarisasi.	54
Gambar 4.9 Perbandingan jumlah data sebelum dan sesudah SMOTE	59
Gambar 4.10. Jumlah data sebelum dan sesudah SMOTEN	61
Gambar 4.11. Confusion matriks best model menggunakan dataset imbalanced.....	63
Gambar 4.12. Confusion matriks best model menggunakan dataset SMOTE	65
Gambar 4.13. Confusion matriks best model dengan dataset SMOTEN	67
Gambar 4.14. Sebaran kelas dataset asli.....	69
Gambar 4.15. Sebaran kelas sesudah SMOTE	69
Gambar 4.16. Sebaran kelas sesudah SMOTEN	69
Gambar 4.17. Grafik perbandingan akurasi.....	70
Gambar 4.18. Perbandingan berdasarkan data split 80:20.....	71
Gambar 4.19. Perbandingan berdasarkan data split 70:30.....	72
Gambar 4.20. Perbandingan akurasi berdasarkan 5 fold	73
Gambar 4.21. Perbandingan akurasi berdasarkan 10 fold	74
Gambar 4.22. Perbandingan akurasi berdasarkan learning rate 0.15.....	75
Gambar 4.23. Perbandingan akurasi berdasarkan learning rate 0.3.....	76
Gambar 4.24. Perbandingan waktu komputasi	78

DAFTAR TABEL

Tabel 2.1. Confusion Matrix	18
Tabel 3.1. Contoh Dataset.....	23
Tabel 3.2. Dataset sebelum cleaning data	26
Tabel 3.3. Sebelum Cleaning Data.....	27
Tabel 3.4. Setelah Cleaning Data.....	28
Tabel 3.5. Contoh dataset sebelum standarisasi.....	29
Tabel 3.6. Contoh dataset setelah standarisasi	30
Tabel 3.7. Dataset contoh sebelum SMOTE.....	32
Tabel 3.8. Dataset contoh setelah implementasi SMOTE	34
Tabel 3.9. Dataset Sebelum SMOTEN	34
Tabel 3.10. Dataset setelah SMOTEN	36
Tabel 3.11. Parameter Booster Tree	38
Tabel 3.12. Contoh dataset untuk implementasi prediksi XGBoost.....	39
Tabel 3.13. Hasil perhitungan gradient dan hessian pada setiap observasi	39
Tabel 3.14. Tampilan Hasil Skenario Pengujian menggunakan Imbalanced Dataset....	46
Tabel 3.15. Tampilan Hasil Skenario Pengujian menggunakan SMOTE Dataset.....	46
Tabel 3.16. Tampilan Hasil Skenario Pengujian menggunakan SMOTEN Dataset.....	46
Tabel 4.1. Hasil skenario menggunakan dataset imbalanced.....	61
Tabel 4.2. Hasil skenario menggunakan dataset SMOTE	64
Tabel 4.3. Hasil skenario menggunakan dataset SMOTEN.....	65
Tabel 4.4. Waktu komputasi teknik resampling.....	77

DAFTAR KODE

Kode 4.1. Modul	48
Kode 4.2. Mengunggah dan membaca dataset.....	51
Kode 4.3. Penyesuaian Data	51
Kode 4.4. Handling Missing Value	52
Kode 4.5. Standarisasi menggunakan z-Score	53
Kode 4.6. Memisahkan fitur dan target	54
Kode 4.7. Mendefinisikan nilai parameter XGBoost.....	55
Kode 4.8. Cross validation.....	56
Kode 4.9. Membaca dan memisahkan fitur	58
Kode 4.10. Strategi resampling SMOTE	58
Kode 4.11. Strategi resampling SMOTEN	60