

KAJIAN *HYBRID SAMPLING* DALAM PENANGANAN *IMBALANCE DATA* PADA ALGORITMA *ENSEMBLE LEARNING* (KASUS RUMAH TANGGA MISKIN DI PROVINSI RIAU)

REKA AGUSTIA ASTARI



**PROGRAM MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Kajian *Hybrid Sampling* dalam Penanganan *Imbalance Data* pada Algoritma *Ensemble Learning* (Kasus Rumah Tangga Miskin di Provinsi Riau)” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada IPB University.

Bogor, Mei 2025

Reka Agustia Astari
G1501221010

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



RINGKASAN

REKA AGUSTIA ASTARI. Kajian *Hybrid Sampling* dalam Penanganan *Imbalance Data* pada Algoritma *Ensemble Learning* (Kasus Rumah Tangga Miskin di Provinsi Riau). Dibimbing oleh **I MADE SUMERTA JAYA** dan **AGUS MOHAMAD SOLEH**.

Penelitian ini mengkaji efektivitas metode *hybrid sampling* dalam menangani ketidakseimbangan data (*imbalanced data*) pada algoritma *ensemble learning*, dengan fokus pada klasifikasi status rumah tangga miskin di Provinsi Riau. Data kemiskinan umumnya memiliki distribusi kelas yang tidak seimbang, di mana proporsi rumah tangga miskin jauh lebih kecil dibandingkan rumah tangga tidak miskin. Kondisi ini menyebabkan algoritma pembelajaran cenderung bias terhadap kelas mayoritas, sehingga akurasi dalam mendeteksi kelas minoritas menjadi rendah.

Untuk mengatasi permasalahan tersebut, penelitian ini mengimplementasikan kombinasi teknik *Synthetic Minority Over-sampling Technique* (SMOTE) dengan *Neighbourhood Cleaning Rule* (NCL) dan *Edited Nearest Neighbour* (ENN) pada dua algoritma *ensemble learning*, yaitu *Double Random Forest* (DRF) dan *Extremely Randomized Trees* (ET). Analisis dilakukan dengan menggunakan data simulasi dan data empiris dari Survei Sosial Ekonomi Nasional (SUSENAS) tahun 2023. Data simulasi dirancang dalam 18 skenario berdasarkan kombinasi variasi rasio ketidakseimbangan data (60:40, 80:20, 95:5) dan tingkat korelasi antar peubah (rendah, tinggi, dan campuran). Metode *hybrid sampling* seperti SMOTE-NCL dan SMOTE-ENN diterapkan untuk menyeimbangkan distribusi kelas, sekaligus meminimalkan risiko *overfitting* yang sering terjadi pada teknik *oversampling* murni.

Evaluasi performa model dilakukan menggunakan metrik *balanced accuracy*, *sensitivity*, dan *G-mean*, dengan pengujian berulang sebanyak 100 kali untuk memastikan stabilitas hasil. Proses pemodelan juga dilengkapi dengan *hyperparameter tuning* dan analisis *Shapley Additive Explanations* (SHAP) untuk mengidentifikasi peubah-peubah yang berkontribusi besar terhadap prediksi status kemiskinan.

Hasil penelitian menunjukkan bahwa kombinasi algoritma DRF dengan metode SMOTE-NCL secara konsisten menghasilkan performa terbaik dibandingkan metode lain. Pada skenario simulasi dengan ketidakseimbangan ekstrem (95:5) dan korelasi rendah, pendekatan ini mencapai nilai *balanced accuracy* sebesar 87,56%, *sensitivity* sebesar 81,85%, dan *G-mean* sebesar 78,59%. Sebaliknya, algoritma ET tanpa penanganan ketidakseimbangan hanya mencapai *balanced accuracy* sekitar 50%, yang mencerminkan adanya bias terhadap kelas mayoritas. Pada data empiris, kombinasi DRF dengan SMOTE-NCL juga menunjukkan hasil terbaik dengan rata-rata *balanced accuracy* sebesar 82,47%, *sensitivity* sebesar 78,92%, dan *G-mean* sebesar 75,30%. Berdasarkan analisis SHAP, peubah yang paling berkontribusi terhadap prediksi status kemiskinan adalah luas lantai, kode wilayah, jumlah anggota rumah tangga, kepemilikan kulkas, rumah, dan tanah, jenis lantai, jenis dinding, fasilitas sanitasi, serta tingkat pendidikan kepala rumah tangga.

Dengan demikian, penelitian ini menyimpulkan bahwa metode *hybrid sampling*, khususnya SMOTE-NCL, dapat secara signifikan meningkatkan akurasi dan ketepatan klasifikasi pada data tidak seimbang. Teknik ini juga mampu merepresentasikan kelas minoritas dengan lebih baik tanpa mengorbankan generalisasi model. Oleh karena itu, pendekatan ini direkomendasikan untuk diterapkan pada kasus serupa di bidang lain yang menghadapi tantangan ketidakseimbangan data, serta sebagai dukungan dalam pengambilan keputusan berbasis data untuk program pengentasan kemiskinan.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Penelitian selanjutnya disarankan untuk mengeksplorasi teknik *hybrid sampling* lainnya dan menguji algoritma pembelajaran mesin yang lebih kompleks guna meningkatkan performa prediktif secara menyeluruh.

Kata kunci: *double random forest, extra trees, hybrid sampling, imbalanced data* dan rumah tangga miskin.

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



SUMMARY

REKA AGUSTIA ASTARI. A Study of Hybrid Sampling for Handling Data Imbalance in Ensemble Learning Algorithms (Case Study of Poor Households in Riau Province). Supervised by **I MADE SUMERTAJAYA** and **AGUS MOHAMAD SOLEH**.

This study investigates the effectiveness of hybrid sampling methods in addressing class imbalance in ensemble learning algorithms, with a specific focus on classifying poor households in Riau Province. Poverty-related data typically exhibit imbalanced class distributions, where the proportion of poor households is significantly lower than that of non-poor households. This imbalance poses a major challenge in classification modeling, as learning algorithms tend to be biased toward the majority class, resulting in low accuracy in detecting the minority class. To mitigate this issue, the study applies a combination of the Synthetic Minority Over-sampling Technique (SMOTE) with Neighbourhood Cleaning Rule (NCL) and Edited Nearest Neighbour (ENN) on two ensemble learning algorithms: Double Random Forest (DRF) and Extremely Randomized Trees (ET). The analysis utilizes both simulated data and empirical data derived from the 2023 National Socioeconomic Survey (SUSENAS). The simulated dataset was designed across 18 different scenarios, considering variations in imbalance ratios (60:40, 80:20, 95:5) and correlation levels among variables (low, high, and mixed). Hybrid sampling methods such as SMOTE-NCL and SMOTE-ENN were used to balance the class distribution while minimizing the risk of overfitting commonly associated with pure oversampling techniques.

Model performance was evaluated using balanced accuracy, sensitivity, and *G-mean* metrics, with 100 iterations of repeated testing to ensure stability of results. The modeling process included hyperparameter tuning and was complemented by Shapley Additive Explanations (SHAP) analysis to identify the most influential predictors in determining household poverty status.

The findings show that the DRF algorithm combined with SMOTE-NCL consistently outperformed other tested methods. Under the simulated scenario with extreme imbalance (95:5) and low correlation, this method achieved a balanced accuracy of 87.56%, sensitivity of 81.85%, and *G-mean* of 78.59%. In contrast, the ET algorithm without imbalance handling performed suboptimally, with a balanced accuracy of approximately 50%, indicating a strong bias toward the majority class. On empirical data, the DRF with SMOTE-NCL also delivered the best results, achieving average balanced accuracy of 82.47%, sensitivity of 78.92%, and *G-mean* of 75.30%.

SHAP analysis revealed that the most influential variables in predicting poverty status included floor area, region code, household size, ownership of a refrigerator, house and land ownership, type of flooring and wall materials, sanitation access, and the highest education level of the household head.

In conclusion, this study demonstrates that hybrid sampling techniques particularly SMOTE-NCL can significantly enhance the performance of classification models in handling imbalanced data. This method not only improves predictive accuracy but also ensures fairer representation of minority classes. Therefore, its application is recommended in similar studies across other domains facing class imbalance

challenges. Furthermore, integrating SHAP analysis is suggested to support data-driven decision-making, particularly in poverty alleviation programs. Future research is encouraged to explore alternative hybrid sampling methods and more complex machine learning algorithms to further enhance predictive accuracy and model interpretability.

Keywords: *double random forest, extra trees, imbalance data, hybrid sampling and poor household.*

@Hak cipta milik IPB University

IPB University





Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

**KAJIAN *HYBRID SAMPLING* DALAM PENANGANAN
IMBALANCE DATA PADA ALGORITMA *ENSEMBLE LEARNING*
(KASUS RUMAH TANGGA MISKIN DI PROVINSI RIAU)**

REKA AGUSTIA ASTARI

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

**PROGRAM MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Penguji Luar Komisi Pembimbing pada Ujian Tesis:
Dr. Indahwati., M.Si



Judul Tesis : Kajian Hybrid *Sampling* dalam Penanganan *Imbalance Data* pada Algoritma *Ensemble Learning* (Kasus Rumah Tangga Miskin di Provinsi Riau)

Nama : Reka Agustia Astari

NIM : G1501221010

Disetujui oleh

Pembimbing 1:

Prof. Dr. Ir. I Made Sumertajaya, M.Si.



Pembimbing 2:

Dr. Agus Mohamad Soleh, S.Si., M.T.



Diketahui oleh

Ketua Program Studi:

Dr. Agus Mohamad Soleh, S.Si., M.T.

NIP 197503151999031004



Dekan Sekolah Sains Data, Matematika, dan Informatika:

Prof. Dr. Ir. Agus Bueno, M.Si., M.Kom.

NIP 196607021993021001



Tanggal Ujian:
06 Maret 2025

Tanggal Lulus:



IPB University

IPB University
— Bogor Indonesia —

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian ini ialah penanganan *imbalance data*, dengan judul “Kajian *Hybrid Sampling* dalam Penanganan *Imbalance Data* pada Algoritma *Ensemble Learning* (Kasus Rumah Tangga Miskin di Provinsi Riau)”.

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada para pembimbing, Prof. Dr. Ir. I Made Sumertajaya, M.Si. dan Dr. Agus Mohamad Soleh, S.Si., M.T. atas bimbingan dan arahnya selama proses penelitian. Ucapan terima kasih juga disampaikan kepada Prof. Dr. Ir. Moh. Yani, M.Eng. selaku moderator seminar, dan Dr. Indahwati, M.Si. selaku penguji luar komisi, atas masukan yang membangun.

Terima kasih kepada Badan Pusat Statistik RI atas dukungan dalam penyediaan data, serta kepada keluarga tercinta atas doa, dukungan, dan kasih sayang yang tiada henti. Penulis juga menyampaikan apresiasi kepada rekan-rekan Kurtosis'22, mahasiswa Pascasarjana Statistika IPB University, yang telah menjadi teman belajar dan berbagi perjuangan, serta kepada rekan-rekan organisasi baik internal maupun eksternal kampus yang telah memberi warna dalam perjalanan studi ini.

Semoga karya ini bermanfaat dan dapat memberikan kontribusi bagi perkembangan ilmu pengetahuan.

Bogor, Mei 2025

Reka Agustia Astari



- 

DAFTAR ISI

DAFTAR TABEL	x
DAFTAR GAMBAR	x
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	4
1.4 Manfaat	4
1.5 Ruang Lingkup	4
II TINJAUAN PUSTAKA	5
2.1 <i>Ensemble Learning</i>	5
2.2 <i>Extremely Randomized Trees (ET)</i> .	6
2.3 <i>Double Random Forest (DRF)</i>	7
2.4 <i>Imbalance Data</i>	8
2.4.1 <i>Neighborhood Cleaning Rule (NCL)</i>	10
2.4.2 <i>Edited Nearest Neighbord (ENN)</i>	10
2.4.3 <i>Synthetic Minority Over-sampling Technique (SMOTE)</i>	10
2.5 <i>Shapley Addictive Explanation (SHAP)</i>	11
2.6 Uji Kebaikan Model	13
2.7 Kemiskinan	14
III METODE	15
3.1 Data	15
3.2 Tahapan Analisis Data	18
IV HASIL DAN PEMBAHASAN	23
4.1 Data Simulasi	23
4.2 Data Empiris	32
4.3 Analisis Kemiskinan di Riau menggunakan SHAP	38
V SIMPULAN DAN SARAN	40
5.1 Simpulan	40
5.2 Saran	40
DAFTAR PUSTAKA	42
RIWAYAT HIDUP	45



DAFTAR TABEL

1	<i>Confusion matrix</i> pada klasifikasi	13
2	Rincian peubah data penelitian	16
3	Skenario pembangkitan data simulasi	16
4	Nilai <i>tuning hyparameter</i> pada DRF dan ET	19
5	Hasil <i>tuning hyperparameter</i> terbaik	23
6	Kinerja model pada ketidakseimbangan rendah.	24
7	Kinerja model pada ketidakseimbangan tinggi.	25
8	Kinerja model pada ketidakseimbangan tinggi.	26
9	Hasil pemodelan data simulasi	29
10	Hasil skenario perbedaan korelasi	31
11	Data ilustratif untuk pembentukan pohon klasifikasi	33
12	Distribusi kelas dan perhitungan <i>split entropy</i> berdasarkan luas lantai.	34
13	Distribusi kelas dan perhitungan <i>split entropy</i> berdasarkan jenis lantai	34
14	Distribusi kelas dan perhitungan <i>split entropy</i> berdasarkan dinding	35
15	Perbandingan rata-rata kinerja model pada data empiris	37

DAFTAR GAMBAR

1	Ilustrasi algoritma ET	7
2	Ilustrasi algoritma DRF	8
3	Skenario ketidakseimbangan ringan	8
4	Skenario ketidakseimbangan sedang	9
5	Skenario ketidakseimbangan tinggi	9
6	Ilustrasi metode SMOTE-NCL dan SMOTE-ENN	11
7	Diagram tahapan penelitian data simulasi	20
8	Diagram tahapan penelitian data empiris	22
9	Perbandingan hasil <i>balanced accuracy</i> data simulasi.	27
10	Perbandingan hasil <i>sensitivity</i> data simulasi	28
11	Perbandingan hasil <i>G-mean</i> data simulasi	29
12	Persentase dan Jumlah Data Kategori.	33
13	Ilustrasi struktur pohon klasifikasi sederhana	36
14	Boxplot rata-rata confusion matrik data empiris	37
15	SHAP Plot: kontribusi terbesar dalam model	38