

**PERBANDINGAN ALGORITMA RESAMPLING PADA DECISION
TREE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS
DATASET PADA IOT DAN CYBER SECURITY DATA**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

MICHAEL SETIAWAN

19.83.0360

Kepada

**PROGRAM SARJANA
PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2023

**PERBANDINGAN ALGORITMA RESAMPLING PADA DECISION
TREE UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS
DATASET PADA IOT DAN CYBER SECURITY DATA**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

MICHAEL SETIAWAN

19.83.0360

Kepada

**PROGRAM SARJANA
PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2023

HALAMAN PERSETUJUAN

SKRIPSI

**PERBANDINGAN ALGORITMA RESAMPLING PADA DECISION TREE
UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS DATASET
PADA IOT DAN CYBER SECURITY DATA**

yang disusun dan diajukan oleh

Michael Setiawan

19.83.0360

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Agustus 2023

Dosen Pembimbing,



Anggit Ferdia Nugraha, S.T., M.Eng
NIK. 190302480

HALAMAN PENGESAHAN
SKRIPSI
PERBANDINGAN ALGORITMA RESAMPLING PADA DECISION TREE
UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS DATASET
PADA IOT DAN CYBER SECURITY DATA

yang disusun dan diajukan oleh

Michael Setiawan

19.83.0360

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Agustus 2023

Susunan Dewan Penguji

Nama Penguji

Ali Mustopa, M.Kom
NIK. 190302192

Andika Agus Slameto, M.Kom
NIK. 190302109

Anggit Ferdita Nouraha, S.T., M.Eng
NIK. 190302480

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Agustus 2023

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Michael Setiawan
NIM : 19.83.0360

Menyatakan bahwa Skripsi dengan judul berikut:

Perbandingan Algoritma Resampling pada Decision Tree untuk Mengatasi Ketidakseimbangan Kelas Dataset pada IoT dan Cyber Security Data

Dosen Pembimbing : Anggit Ferdita Nugraha, S.T., M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 21 Agustus 2023

Yang Menyatakan,



Michael Setiawan

HALAMAN PERSEMBAHAN

Puji syukur atas upaya dan dedikasi yang telah memungkinkan penulis menyelesaikan skripsi ini. Dengan rasa hormat, saya ingin menyampaikan hasil penelitian ini sebagai ungkapan terima kasih kepada orang tua, dosen pembimbing, dan teman-teman yang selalu memberikan dukungan dan kasih sayang tanpa henti. Kontribusi Anda telah membantu saya dalam mencapai tujuan hidup dan menjalani kehidupan dengan penuh semangat.



KATA PENGANTAR

Dengan rasa terima kasih yang mendalam, penulis ingin menghaturkan apresiasi atas kemampuan dan bimbingan yang telah memungkinkan penyelesaian Skripsi dengan judul: "Perbandingan Algoritma Resampling pada Decision Tree untuk Mengatasi Ketidakseimbangan Kelas Dataset pada IoT dan Cyber Security Data" Serta, penghormatan kepada individu yang telah memimpin kita dari masa lalu menuju masa depan yang terang. Sebagai bentuk ekspresi rasa puji dan syukur atas penyelesaian penulisan Skripsi ini, penulis ingin mengucapkan terima kasih kepada:

1. Bapak Anggit Ferdita Nugraha.ST., M.Eng, sebagai pembimbing yang telah memberikan arahan dan panduan sehingga penulis mampu menyelesaikan skripsi ini.
2. Orang tua saya yang senantiasa memberikan dukungan yang tiada hentinya kepada saya.
3. Teman saya yang selalu bertukar saran dan ide.
4. Yang terakhir saya ucapkan banyak terima kasih kepada diri saya sendiri yang telah mendorong diri ini yang penuh dengan kemageran untuk dapat menyelesaikan skripsi ini pada tepat waktu.

Yogyakarta, 21 Agustus 2023

Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL	ix
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	xi
DAFTAR LAMBANG DAN SINGKATAN	xii
DAFTAR ISTILAH	xiii
INTISARI	xiv
ABSTRACT.....	xv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	2
1.5 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA	4
2.1 Studi Literatur	4
2.2 Dasar Teori	9
2.2.1 Imbalance Data	9
2.2.2 Artificial Intelligence (AI)	9
2.2.3 Machine Learning (ML)	10
2.2.3.1 Supervised Learning	10
2.2.3.2 Unsupervised Learning	11
2.2.3.3 Semi Supervised Learning	11
2.2.3.4 Reinforcement Learning	11
2.2.4 Klasifikasi	12
2.2.5 Decision Tree (DT).....	13
2.2.6 SMOTE.....	14
2.2.7 ADASYN.....	14
2.2.8 Borderline SMOTE.....	14

2.2.9	SVM SMOTE	14
2.2.10	SMOTENC	15
2.2.11	Random Over Sampler (ROS)	15
2.2.12	Random Under Sampler (RUS)	15
2.2.13	Edited Nearest Neighbours (ENN)	15
2.2.14	Cluster Centroids	16
2.2.15	Near Miss	16
2.2.16	Neighbourhood Cleaning Rule (NCL)	16
2.2.17	One-Sided Selection (OSS)	17
2.2.18	Evaluasi	17
2.2.18.1	Akurasi	18
2.2.18.2	Presisi	18
2.2.18.3	Recall	18
2.2.18.4	F1 Score	18
BAB III METODE PENELITIAN		20
3.1	Alat dan Bahan	20
3.2	Alur Penelitian	20
3.2.1	Data Acquisition	21
3.2.2	Preprocessing	23
3.2.3	Feature Enggining	23
3.2.4	Modeling dan Analysis	23
3.2.5	Evaluasi	24
BAB IV HASIL DAN PEMBAHASAN		25
4.1	Implementasi	25
4.1.1	Dataset	25
4.1.2	Preprocessing	26
4.1.3	Feature Enggining	34
4.1.4	Modeling	34
4.1.5	Evaluasi	37
BAB V PENUTUP		59
5.1	Kesimpulan	59
5.2	Saran	60
REFERENSI		61
LAMPIRAN		67

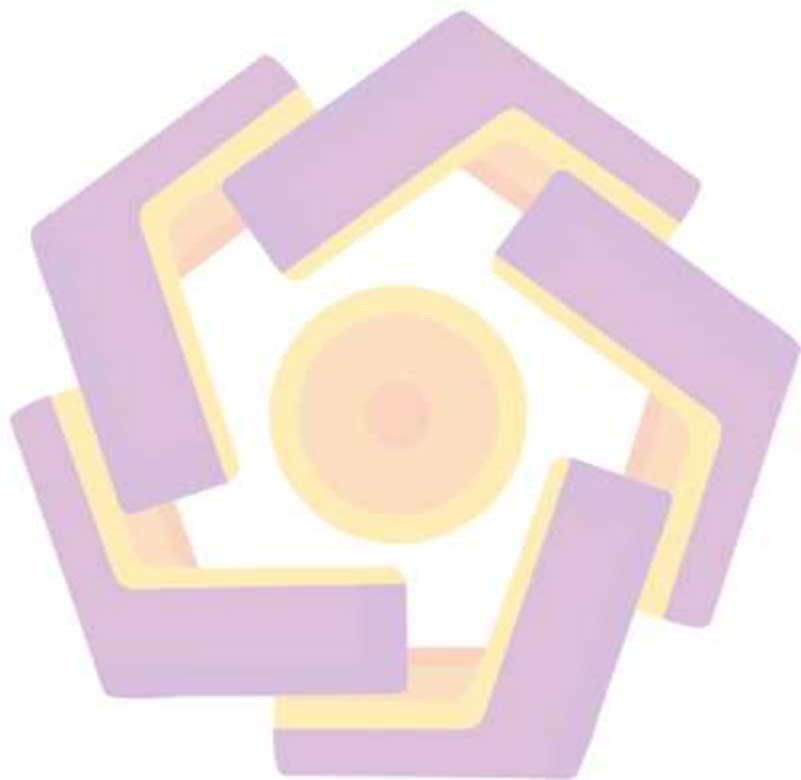
DAFTAR TABEL

Tabel 2.1. Keaslian Penelitian	6
Tabel 3.1. Jumlah Tipe Atribut Pada Dataset	21
Tabel 3.2. Imbalance Ration dan Jumlah Instance	22
Tabel 3.3. Teknik Resampling	23
Tabel 4.1. Hasil Resampling Dataset 1	38
Tabel 4.2. Visualisasi Hasil Resampling Dataset 1	38
Tabel 4.3. Hasil Resampling Dataset 2	40
Tabel 4.4. Visualisasi Hasil Resampling Dataset 2	40
Tabel 4.5. Hasil Resampling Dataset 3	41
Tabel 4.6. Visualisasi Hasil Resampling Dataset 3	42
Tabel 4.7. Hasil Resampling Dataset 4	43
Tabel 4.8. Visualisasi Hasil Resampling Dataset 4	44
Tabel 4.9. Hasil Resampling Dataset 5	45
Tabel 4.10. Visualisasi Hasil Resampling Dataset 5	46
Tabel 4.11. Hasil Resampling Dataset 6	47
Tabel 4.12. Visualisasi Hasil Resampling Dataset 6	48
Tabel 4.13. Hasil Resampling Dataset 7	49
Tabel 4.14. Visualisasi Hasil Resampling Dataset 7	50
Tabel 4.15. Hasil Resampling Dataset 8	51
Tabel 4.16. Visualisasi Hasil Resampling Dataset 8	51
Tabel 4.17. Performa Decision Tree pada Dataset 1	53
Tabel 4.18. Performa Decision Tree pada Dataset 2	54
Tabel 4.19. Performa Decision Tree pada Dataset 3	54
Tabel 4.20. Performa Decision Tree pada Dataset 4	55
Tabel 4.21. Performa Decision Tree pada Dataset 5	56
Tabel 4.22. Performa Decision Tree pada Dataset 6	56
Tabel 4.23. Performa Decision Tree pada Dataset 7	57
Tabel 4.24. Performa Decision Tree pada Dataset 8	58

DAFTAR GAMBAR

Gambar 2.1. Sub Bagian Artificial Intelligence	9
Gambar 2.2. Kategori Machine Learning	10
Gambar 2.3. Tipe Klasifikasi	12
Gambar 2.4. Binaryclass Classification	12
Gambar 2.5. Multiclass Classification	13
Gambar 2.6. Struktur Decision Tree	13
Gambar 2.7. Confusion Matrix	17
Gambar 3.1. Alur Penelitian	20
Gambar 4.1. Dataset 1	25
Gambar 4.2. Dataset 2	25
Gambar 4.3. Dataset 3	25
Gambar 4.4. Dataset 4	25
Gambar 4.5. Dataset 5	26
Gambar 4.6. Dataset 6	26
Gambar 4.7. Dataset 7	26
Gambar 4.8. Dataset 8	26
Gambar 4.9. Handling Kolom Tidak Dipakai	27
Gambar 4.10. Handling Data NaN, NULL dan Duplicate	27
Gambar 4.11. Spliting X dan y	28
Gambar 4.12. Array Untuk Menampung Attribute	28
Gambar 4.13. Loop Pemilahan Attribute	29
Gambar 4.14. Function Manual Scalling	29
Gambar 4.15. Merubah Tipe Data Attribute Continue ke float64	30
Gambar 4.16. Manual Scalling	30
Gambar 4.17. Handling INF value	31
Gambar 4.18. Library Scaling	31
Gambar 4.19. Menggabungkan Atribut	32
Gambar 4.20. Sebelum Label Encoding	32
Gambar 4.21. Eksekusi Label Encoding	33
Gambar 4.22. Sesudah Label Encoding	33
Gambar 4.23. Mendefinisi Library Resampling dan Implementasi	34
Gambar 4.24. Mendefinisi Statified K-Fold dan Decision Tree	34
Gambar 4.25. List Score Hasil Klasifikasi	35
Gambar 4.26. Klasifikasi Dataset	36
Gambar 4.27. Evaluasi Klasifikasi	36
Gambar 4.28. Perbandingan kelas sebelum dan sesudah resampling	37

DAFTAR LAMPIRAN

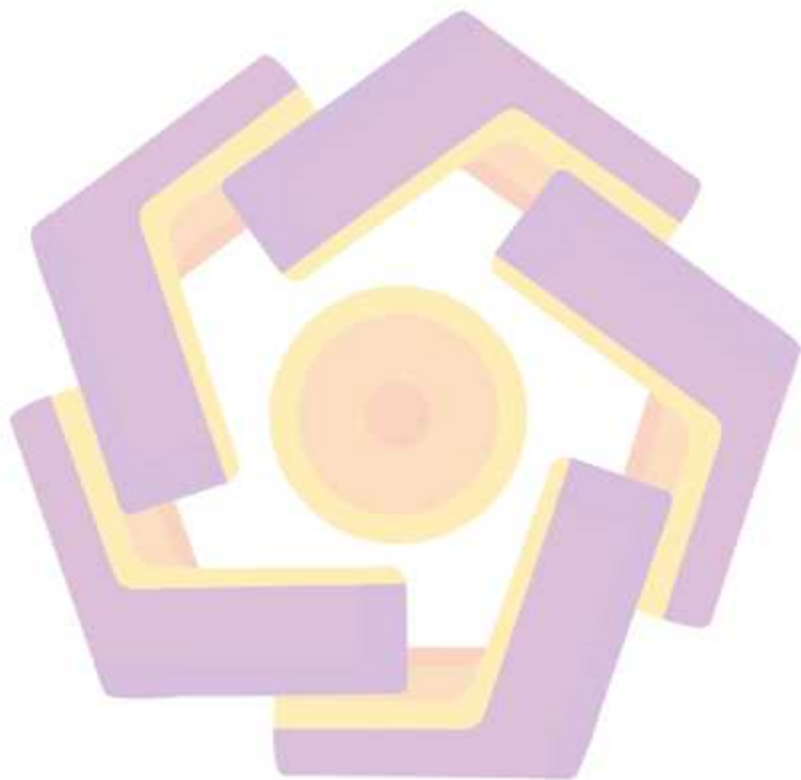


DAFTAR LAMBANG DAN SINGKATAN



AI	Artificial Intelligence
ML	Machine Learning
DT	Decision Tree
SMOTE	Synthetic Minority Over-sampling Technique
ADASYN	Adaptive Synthetic Sampling
SVM	Support Vector Machine
SMOTENC	Synthetic Minority Over-sampling Technique Nominal Continuous
ROS	Random Over-sampling
RUS	Random Under-sampling
ENN	Edited Nearest Neighbors
NCL	Neighborhood Cleaning Rule
OSS	One-Sided Selection
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
INF	Infinity
NaN	Not a Number

DAFTAR ISTILAH



INTISARI

Machine Learning saat ini sudah canggih sehingga bisa membantu pekerjaan manusia, namun Imbalance dataset menjadi salah satu tantangan dalam melakukan klasifikasi pada machine learning saat ini, hal ini dikarenakan data yang berasal dari dunia nyata sering condong ke salah satu arah, contoh dataset pendeteksi api, yang dimana data yang digenerate oleh sensor akan lebih banyak menunjukkan bahwa tidak ada api dibanding dengan adanya api, hal ini menjadikan masalah baru dalam klasifikasi terutama jika jumlah yang diproses berjumlah besar. Saat ini sudah banyak teknik oversampling maupun teknik undersampling yang bisa digunakan untuk menyeimbangkan data, namun teknik tersebut memiliki cara kerjanya masing masing dan belum tentu bisa diaplikasikan kedalam semua jenis dataset. Penelitian ini bertujuan untuk mencari tahu kecocokan teknik resampling terhadap jenis dataset, metodologi yang digunakan adalah eksperimen dengan cara menerapkan 6 teknik oversampling dan 6 teknik undersampling kepada 8 dataset dan akan diklasifikasi menggunakan model Decision Tree. Didapatkan hasil bahwa terdapat keterbatasan beberapa teknik dalam implementasinya untuk menyeimbangkan kelas data, dan hasil kinerja tertinggi yang didapat dalam penelitian ini adalah 100% pada semua teknik yang diterapkan, namun di mungkinkan terjadil overfitting pada dataset tersebut. Dari penelitian ini dapat diketahui bahwa tidak semua teknik resampling bisa diterapkan ke semua dataset, dan dataset yang berukuran sangat besar bisa memakan waktu yang lama sehingga masih belum di dapatkan hasil dari pengaplikasian teknik resampling.

Kata kunci: Machine Learning, Decision Tree, Oversampling, Undersampling, Imbalance.

ABSTRACT

In the present time, Machine Learning has undergone substantial advancements, providing significant assistance to various human tasks. However, a persistent challenge remains in the form of imbalanced datasets within machine learning classification. This challenge arises due to the inherent bias often present in real-world data, leading to complexities in classification tasks, especially when dealing with extensive datasets. While oversampling and undersampling techniques have been developed to tackle this issue, their applicability varies depending on the characteristics of the datasets. This research's objective is to assess the appropriateness of resampling techniques across a range of diverse datasets. Through a series of experiments, six oversampling and six undersampling techniques are applied to eight distinct datasets, followed by classification using a Decision Tree model. The findings underscore the limitations of specific techniques in effectively rebalancing data distribution. Notably, the study achieves a peak performance of 100%, though this result raises concerns about potential overfitting. The research emphasizes the importance of a discerning approach when choosing resampling techniques based on the unique attributes of each dataset, while also acknowledging the time-related challenges posed by handling substantial datasets, which can impact the timely application of resampling methods.

Keyword: Machine Learning, Decision Tree, Oversampling, Undersampling, Imbalance.