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 Ferdita 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	Tanda Tangan
Ali Mustopa, M.Kom **NIK. 190302192**	
Andika Agus Slameto, M.Kom **NIK. 190302109**	
Anggit Ferdita Nugraha, S.T., M.Eng **NIK. 190302480**	

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.

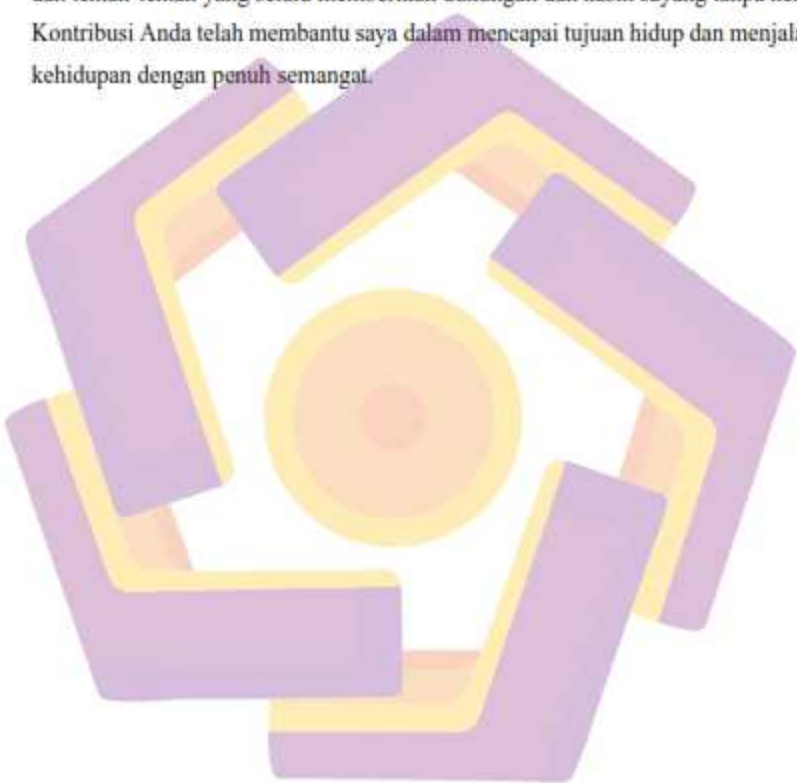
Yogyakarta, 21 Agustus 2023

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

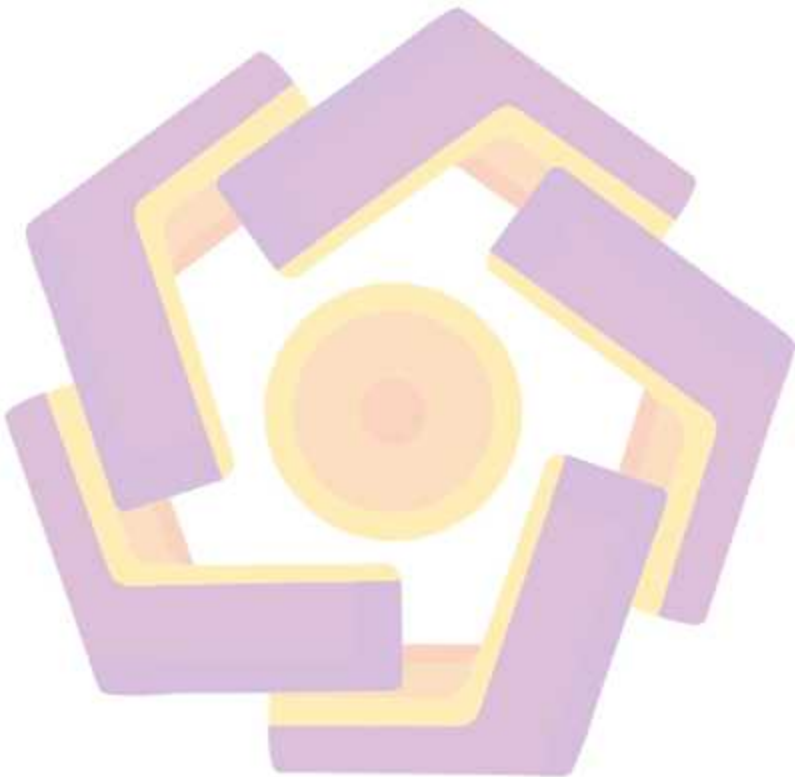
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

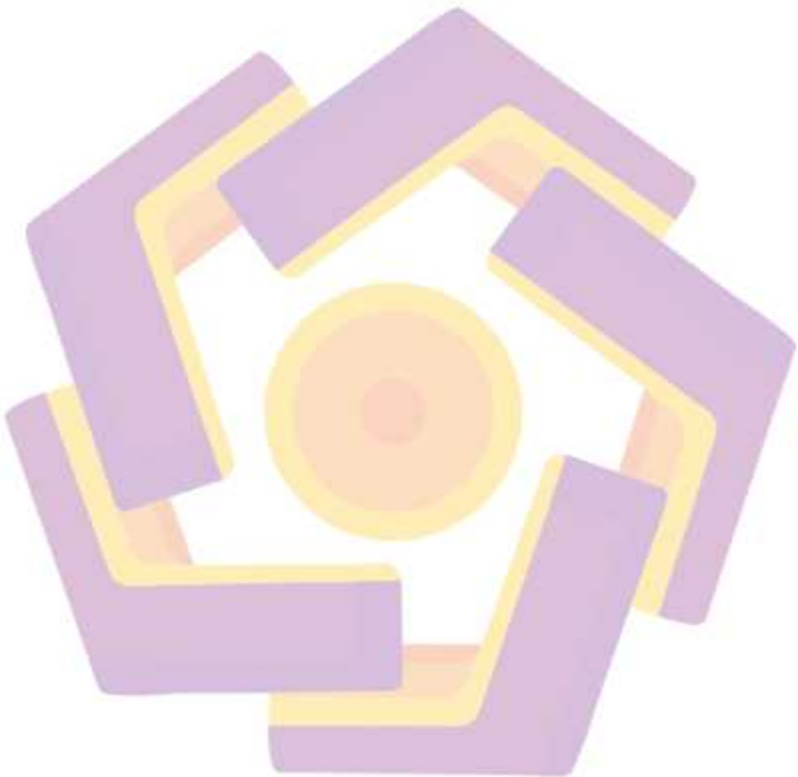
DAFTAR LAMPIRAN



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 menunujukan 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 berurukuran 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.