SKRIPSI

ANALISIS PENGGUNAAN TEKNIK OVERSAMPLING PADA EXTREME GRADIENT BOOSTING (XGBOOST) UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT JANTUNG

MUHAMMAD RAFLI AULIA ROJANI LUTFI
NPM: 20081010061

DOSEN PEMBIMBING
Achmad Junaidi, S.Kom., M.Kom.
Agung Mustika Rizki, S.Kom., M.Kom.

KEMENTRIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2024

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

ANALISIS PENGGUNAAN TEKNIK OVERSAMPLING PADA EXTREME GRADIENT BOOSTING (XGBOOST) UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT JANTUNG

Oleh:
MUHAMMAD RAFLI AULIA ROJANI LUTFI
NPM. 20081010061

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada Tanggal 2 September 2024

Menyetujui,

Achmad Junaidi, S.Kom., M.Kom. ..... **(Pembimbing I)**
NPT. 3 7811 04 0199 1

Agung Mustika Rizki, S.Kom., M.Kom. ..... **(Pembimbing II)**
NIP. 19930725 202203 1008

Dr. Ir. I Gede Susrama Mas , ST. MT. IPU ..... **(Ketua Penguji)**
NIP. 19700619 2021211 009

Retno Mumpuni, S.Kom., M.Sc. ..... **(Penguji I)**
NPT. 172198 70 716054

Mengetahui,
Dekan Fakultas Ilmu Komputer

Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

Halaman ini sengaja dikosongkan

LEMBAR PERSETUJUAN

ANALISIS PENGGUNAAN TEKNIK OVERSAMPLING PADA EXTREME GRADIENT BOOSTING (XGBOOST) UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT JANTUNG

Oleh:

MUHAMMAD RAFLI AULIA ROJANI LUTFI

NPM. 20081010061

Menyetujui,

Koordinator Program Studi Informatika

Fakultas Ilmu Komputer

Fetty Tri Anggraeny, S.Kom., M.Kom.

NIP. 19820211 2021212 005

SURAT PERNYATAAN ORISINALITAS

Yang bertandatangan di bawah ini:

Nama Mahasiswa	:	Muhammad Rafli Aulia Rojani Lutfi
Program Studi	:	Informatika
Dosen Pembimbing	:	1. Achmad Junaidi, S.Kom., M.Kom.
		2. Agung Mustika Rizki, S.Kom., M.Kom.

dengan ini menyatakan bahwa isi sebagian maupun keseluruhan disertasi dengan judul:

ANALISIS PENGGUNAAN TEKNIK OVERSAMPLING PADA EXTREME GRADIENT BOOSTING (XGBOOST) UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA KLASIFIKASI PENYAKIT JANTUNG

adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diizinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri. Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, 10 September 2024
Yang Membuat Pernyataan,

M. RAFLI AULIA ROJANI LUTFI
NPM. 20081010061

Halaman ini sengaja dikosongkan

ABSTRAK

Nama Mahasiswa / NPM	:	Muhammad Rafli Aulia Rojani Lutfi / 20081010061
Judul Skripsi	:	Analisis Penggunaan Teknik Oversampling Pada Extreme Gradient Boosting (XGBoost) Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Penyakit Jantung
Dosen Pembimbing	:	1. Achmad Junaidi, S.Kom., M.Kom. 2. Agung Mustika Rizki, S.Kom., M.Kom.

Penyakit jantung merupakan ancaman serius bagi kesehatan global. Deteksi dini menjadi kunci untuk meningkatkan angka keberlangsungan hidup. Namun, upaya membangun model prediksi seringkali terhambat oleh ketidakseimbangan data, di mana jumlah individu sehat jauh lebih banyak. Ketidakseimbangan ini berpotensi membuat model lebih condong pada kelas mayoritas (sehat) dan mengabaikan kelas minoritas (sakit), sehingga mengurangi akurasi dalam mendeteksi penyakit jantung.

Penelitian ini bertujuan untuk mengatasi masalah ketidakseimbangan kelas dengan menerapkan teknik *oversampling* pada algoritma XGBoost. Teknik *oversampling* bekerja dengan meningkatkan jumlah data pada kelas minoritas sehingga keseimbangan data dapat tercapai. Dalam penelitian ini, tiga teknik *oversampling* yang digunakan, yaitu ROS, SMOTE, dan ADASYN, diuji coba. Selain itu, dilakukan penyetelan hyperparameter pada algoritma XGBoost untuk mendapatkan kinerja model yang optimal.

Hasil penelitian menunjukkan bahwa model XGBoost tanpa *oversampling* yang menggunakan parameter bawaan memiliki akurasi tinggi yaitu di atas 0,90, namun kurang baik dalam mengklasifikasikan kelas minoritas (individu sakit), dibuktikan dengan nilai *g-mean* yang rendah yaitu di bawah 0,50. Sebaliknya, model XGBoost dengan teknik *oversampling* pada rasio *sampling* dan *hyperparameter* yang optimal mampu meningkatkan kinerja model dalam mengklasifikasikan kelas minoritas. Dari tiga teknik *oversampling*, ROS memberikan rata-rata nilai *g-mean* tertinggi sebesar 0,80. Namun, model ini rentan terhadap *overfitting* pada rasio sampling yang lebih besar. Berbeda dengan SMOTE dan ADASYN, meskipun kedua model juga rentan terhadap *overfitting* pada rasio sampling di atas 0,1, namun tingkat kerentanannya jauh lebih rendah dan cenderung lebih stabil.

Kata kunci: Penyakit Jantung, Ketidakseimbangan Kelas, Teknik *Oversampling*, XGBoost

Halaman ini sengaja dikosongkan

ABSTRACT

Student Name / NPM	:	Muhammad Rafli Aulia Rojani Lutfi / 20081010061
Thesis Title	:	Analysis of Oversampling Techniques on Extreme Gradient Boosting (XGBoost) for Handling Class Imbalance in Heart Disease Classification
Advisor	:	1. Achmad Junaidi, S.Kom., M.Kom. 2. Agung Mustika Rizki, S.Kom., M.Kom.

Heart disease is a serious threat to global health. Early detection is key to improving survival rates. However, building predictive models are often hampered by data imbalance, where the number of healthy individuals far exceeds the number of diseased individuals. This imbalance can cause the model to be biased towards the majority class (healthy) and neglect the minority class (diseased), reducing its accuracy in detecting heart disease.

This research aims to address the problem of class imbalance by applying oversampling techniques to the XGBoost algorithm. The oversampling technique works by increasing the amount of data in the minority class so that data balance can be achieved. In this study, three oversampling techniques, namely ROS, SMOTE, and ADASYN, were tested. In addition, hyperparameter tuning of the XGBoost algorithm was performed to obtain optimal model performance.

The results show that the XGBoost model without oversampling using default parameters has a high accuracy of above 0.90, but is not good at classifying minority classes (sick individuals), as evidenced by the low g-mean value of below 0.50. In contrast, the XGBoost model with oversampling techniques at optimal sampling ratios and hyperparameters is able to improve the performance of the model in classifying the minority class. Of the three oversampling techniques, ROS provides the highest average g-mean value of 0.80. However, the model is prone to overfitting at larger sampling ratios. In contrast to SMOTE and ADASYN, although both models are also susceptible to overfitting at sampling ratios above 0.1, the degree of susceptibility is much lower and tends to be more stable.

Keywords: Heart Disease, Imbalance Class, Oversampling Techniques, XGBoost

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Segala puji dan Syukur kehadirat Allah SWT yang telah melimpahkan rahmat-Nya, sehingga penulis dapat menyelesaikan penelitian skripsi beserta laporan hasil skripsi yang berjudul "Analisis Penggunaan Teknik Oversampling Pada Extreme Gradient Boosting (Xgboost) Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Penyakit Jantung" dengan baik.

Penyelesaian laporan skripsi ini tidak lepas dari dukungan dan bantuan dari berbagai pihak. Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada semua pihak yang telah berkontribusi dalam kelancaran penelitian ini, mulai dari tahap perencanaan hingga penyusunan laporan akhir. Secara khusus, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Ir. Akhmad Fauzi, MMT. Selaku Rektor Univeristas Pembangunan Nasional "Veteran" Jawa Timur.
2. Ibu Dr. Ir. Novirina Hendrasarie, MT. selaku Dekan Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
3. Ibu Fetty Tri Anggraeny, S.Kom., M.Kom. selaku Koordinator Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
4. Bapak Achmad Junaidi, S.Kom., M.Kom. selaku Dosen Pembimbing I, yang telah membimbing penulis selama pengerjaan skripsi ini.
5. Bapak Agung Mustika Rizki, S.Kom., M.Kom. selaku Dosen Pembimbing II, yang telah membimbing penulis selama pengerjaan skripsi ini.
6. Seluruh Dosen dan Staff Program Studi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional "Veteran" Jawa Timur atas segala bantuan yang diberikan untuk membantu penulis dalam penyusunan skripsi ini.
7. Ibunda dan Ayahanda tercinta yang tidak pernah putus mendoakan dan memberikan dukungan moril serta meteri kepada penulis selama menempuh studi.
8. Teman-teman "Setunggal Riset" yang telah memberikan dukungan selama masa perkuliahan dan penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih penuh kekurangan, oleh karena itu saran dan kritik yang membangun sangat penulis harapkan. Besar harapan penulis bahwa dengan adanya penelitian ini dapat memberikan manfaat kepada para pembaca. Amin yaa Robbal ‘Alamin.

Surabaya, 25 Agustus 2024

Penulis

DAFTAR ISI

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

DAFTAR TABEL

Halaman ini sengaja dikosongkan

DAFTAR NOTASI

X^2	:	nilai *Chi-Square*
f_o	:	Frekuensi observasi
f_e	:	Frekuensi ekspetasi
b_i	:	jumlah nilai baris ke-i
k_j	:	jumlah nilai kolom ke-j
N	:	total seluruh nilai observasi
d(x, y)	:	jarak data x dan y
n	:	jumlah dimensi (atribut)
x_i	:	atribut ke i dari data x
y_i	:	atribut ke i dari data y
x_{syn}	:	data sintetis
x_{knn}	:	tetangga terdekat dari xi
γ	:	bilangan desimal acak antara 0 dan 1
d	:	*degree of class imbalance*
m_s	:	jumlah kelas minoritas
m_l	:	jumlah kelas mayoritas
β	:	tingkat keseimbangan data sintetis
Δ_i	:	jumlah data mayoritas pada K tetangga terdekat dari data minoritas
x_i	:	titik data kelas minoritas asli
x_{zi}	:	titik data tetangga terdekat
λ	:	bilangan desimal acak antara 0 dan 1
$\hat{y}_i$	:	jumlah daun pada pohon keputusan
T	:	parameter regularisasi ke-1
t	:	iterasi ke-t
$\hat{y}_i^{(t-1)}$	:	prediksi model pada iterasi sebelumnya
$f_t(x_i)$	:	prediksi pohon keputusan ke-t untuk sampel data ke-i

Halaman ini sengaja dikosongkan