ALGORITMA XGBOOST UNTUK MENINGKATKAN KINERJA MODEL MACHINE LEARNING PADA MULTICLASS IMBALANCED DATASET

SKRIPSI

disusun oleh :

Zulfikar Murakabiman

18.11.2474

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2022

ALGORITMA XGBOOST UNTUK MENINGKATKAN KINERJA MODEL MACHINE LEARNING PADA MULTICLASS IMBALANCED DATASET

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika

disusun oleh :

Zulfikar Murakabiman

18.11.2474

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2022

PERSETUJUAN

SKRIPSI

ALGORITMA XGBOOST UNTUK MENINGKATKAN KINERJA MODEL MACHINE LEARNING PADA MULTICLASS IMBALANCED DATASET

yang dipersiapkan dan disusun oleh

Zulfikar Murakabiman

18.11.2474

telah disetujui oleh Dosen Pembimbing Skripsi

pada tanggal 9 Februari 2022

Dosen Pembimbing,

Yoga Pristyanto, S.Kom., M.Eng.

NIK. 190302412

PENGESAHAN

SKRIPSI

ALGORITMA XGBOOST UNTUK MENINGKATKAN KINERJA MODEL MACHINE LEARNING PADA MULTICLASS IMBALANCED DATASET

yang dipersiapkan dan disusun oleh
Zulfikar Murakabiman

18.11.2474

telah dipertahankan di depan Dewan Penguji
pada tanggal 17 Februari 2022

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Arif Akbarul Huda, S.Si, M.Eng **NIK. 190302287**	_____
Uyock Anggoro Saputro, M.Kom **NIK. 190302419**	_____
Yoga Pristyanto, S.Kom., M.Eng **NIK. 190302412**	_____

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 April 2022

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, M.Kom
NIK. 190302096

PERNYATAAN

Saya yang bertandatangan dibawah ini menyatakan bahwa, skripsi ini merupakan karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

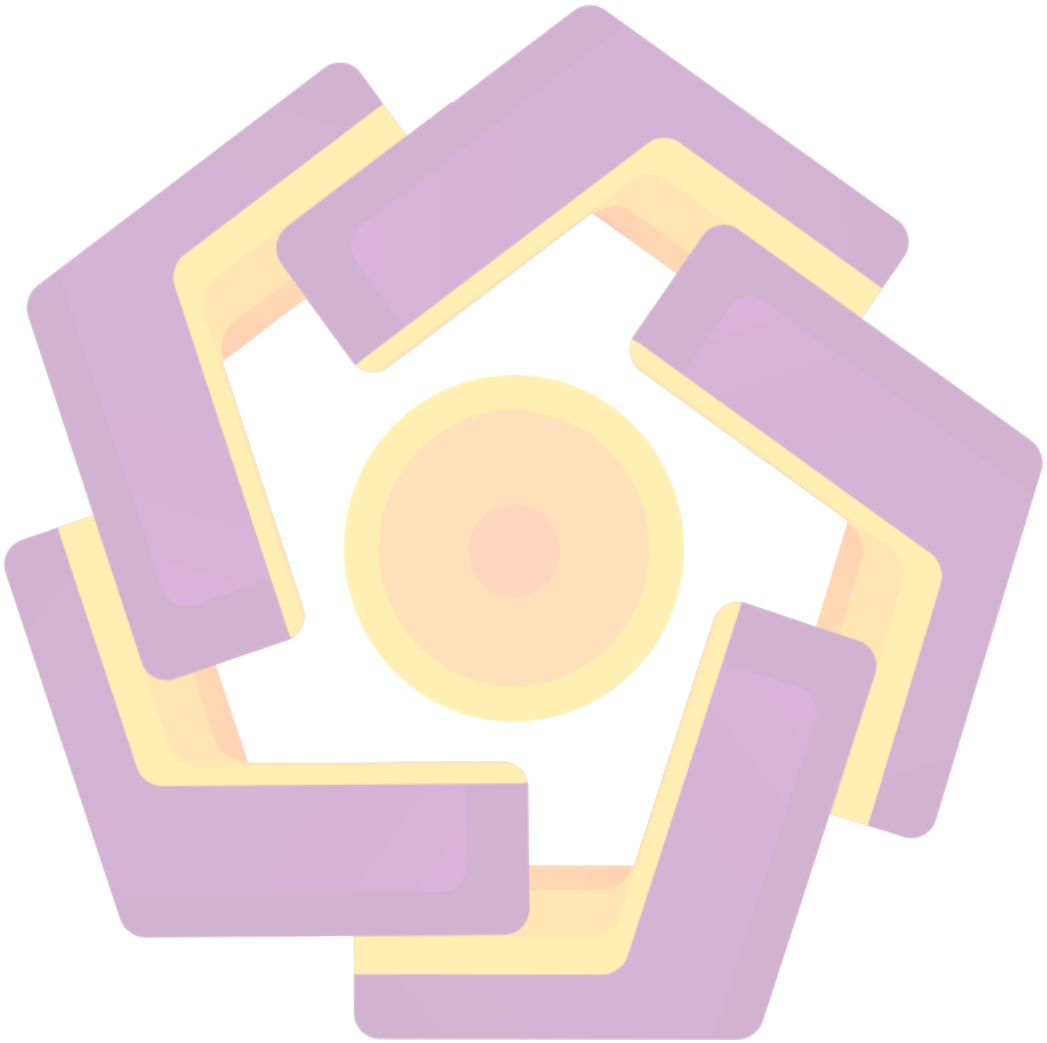
Yogyakarta, 19 April 2022

Zulfikar Murakabiman
NIM. 18.11.2474

MOTTO

Luck is what happens when preparation meets oppoturnity

(Seneca)



PERSEMBAHAN

Alhamdulillah penulis panjatkan pujisyukur kepada Allah SWT atas segala rahmat, taufiq, serta hidayah-Nya, sehingga diberi kesempatan untuk dapat menyelesaikan skripsi ini dengan sebaik-baiknya dengan segala kekurangan penulis. Segala syukur penulis ucapkan kepada-Mu karena telah menghadirkan mereka yang memberikan semangat dan doa disaat menjalani proses pembuatan skripsi ini. Dengan segala kerendahan hati saya persembahkan skripsi ini kepada semua orang yang terlibat dan mendukung saya dalam pengerjaan skripsi ini.

Saya ucapkan terima kasih yang sebesar-besarnya untuk kalian semua. Mohon maaf jika ada salah kata atau perbuatan baik yang disengaja maupun tidak disengaja selama ini. Sukses untuk kalian semua, semoga Allah SWT memberikan rahmat dan hidayah-Nya kepada kita semua. Dan semoga skripsi ini dapat bermanfaat dan berguna untuk kemajuan ilmu pengetahuan kedepannya.

KATA PENGANTAR

Alhamdulillah penulis panjatkan puji syukur kepadaAllah SWT atas segala rahmat, taufiq, serta hidayah-Nya kepada penulis sehingga dapat menyelesaikan skripsi yang berjudul "Algoritma XGBoost Untuk Meningkatkan Kinerja Model Machine Learning pada Multiclass Imbalanced Dataset".

Selama prosespengerjaan skripsi ini penulis menyadari bahwa dalam proses penulisan skripsi ini banyak mengalami kendala, namun berkat bantuan, bimbingan, kerjasama dari berbagai pihak dan berjkah dari Allah SWT sehingga kendala-kendala tersebut bisa diatasi. Selanjutnya ucapa terima kasih penulis sampaikan kepada :

1. Bapak Prof. Dr. M. Suyanto, M.M selaku Rektor Universitas Amikom Yogyakarta.
2. Bpk. Yoga Pristyanto, S.Kom, M.Eng selaku dosen pembimbing yang telah memberikan banyak masukkan yang membantu membimbing dalam menyelesaikan skripsi ini.
3. Kedua orang tua dan keluarga yang selalu memberikan doa, dukungan dan semangat.
4. Serta semua pihak yang tidak bisa penulis sebutkan satu-persatu yang telah membantu dalam penyusunan skripsi ini.

Penulis Menyadari bahwa masih banyak terdapat kekurangan-kekurangan dalam mengerjakan skripsi ini, sehingga penulis mengharapkan adanya saran dan kritik yang membangun demi kesempurnaan skripsi ini.

Yogyakarta, 10 Februari 2022

Zulfikar Murakabiman

NIM. 18.11.2474

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN JUDUL ii
PERSETUJUAN iii
PENGESAHAN iv
PERNYATAAN v
MOTTO vi
PERSEMBAHAN vii
KATA PENGANTAR viii
DAFTAR ISI ix
DAFTAR GAMBAR xi
DAFTAR TABEL xii
INTISARI xiii
ABSTRACT xiv
BAB I 1
PENDAHULUAN 1
1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 6
1.3 Batasan Penelitian 6
1.4 Tujuan Penelitian 7
1.5 Manfaat Penelitian 7
1.6 Metode Penelitian 8
1.6.1 Metode Pengumpulan Data. 8
1.6.2 Metode Klasifikasi dan Penanganan Ketidakseimbangan Kelas. 8
1.6.3 Metode Evaluasi. 8
1.7 Sistematika Penulisan 9
BAB II 11
TINJUAN PUSTAKA DAN LANDASAN TEORI 11
2.1 Tinjauan Pustaka 11
2.2 Landasan Teori 22

2.2.1 Data Mining 22
2.2.2 Boosting 22
2.2.3 XGBoost 24
2.2.4 Data Preprocessing 26
2.2.5 Evaluasi Model 27
BAB III 31
METODOLOGI PENELITIAN 31
3.1 Gambaran Umum 31
3.2 Alat dan Bahan 31
3.2.1 Alat 31
3.2.2 Bahan 32
3.3 Jalannya Penelitian 40
3.4 Pra Pemrosesan Data 41
3.5 Klasifikasi dan Penanganan Ketidakseimbangan Kelas 41
3.6 Evaluasi 43
BAB IV 45
HASIL DAN PEMBAHASAN 45
4.1 Dataset 45
4.2 Pra Pemrosesan Data 47
4.3 Implementasi Algoritma Klasifikasi 48
4.3.1 Nilai Balanced Accuracy 48
4.3.2 Nilai G-Mean 50
4.3.3 Nilai MAUC 52
4.3.4 Nilai Sensitivity 54
4.3.5 Nilai Specificity 56
BAB V 59
KESIMPULAN DAN SARAN 59
5.1 Kesimpulan 59
5.2 Saran 59
DAFTAR PUSTAKA 61

DAFTAR GAMBAR

Gambar 3.3.1 Alur Penelitian 40
Gambar 3.5.1 Gambaran Model XGBoost..... 42
Gambar 4.1.1 Grafik Distribusi Kelas Dataset..... 47
Gambar 4.3.1 Diagram Nilai Balanced Accuracy..... 49
Gambar 4.3.2 Diagram Nilai G-Mean..... 51
Gambar 4.3.3 Diagram Nilai MAUC..... 53
Gambar 4.3.4 Diagram Nilai Sensitivity..... 55
Gambar 4.3.5 Diagram Nilai Specificity..... 57

DAFTAR TABEL

Tabel 2.2.1 Contoh Missing Value 27
Tabel 2.2.2 Confusion Matrix 28
Tabel 3.2.1 Karakteristik Dataset 32
Tabel 3.2.2 car 34
Tabel 3.2.3 contraceptive 34
Tabel 3.2.4 glass 35
Tabel 3.2.5 hayes-roth 35
Tabel 3.2.6 new-thyroid 36
Tabel 3.2.7 pageblocks 37
Tabel 3.2.8 winequality-red 37
Tabel 3.2.9 winequality-white 38
Tabel 3.2.10 yeast 39
Tabel 3.6.1 XGBoost Confusion Matrix 43
Tabel 4.2.1 Hasil Dataset yang Menunjukan Nilai Missing Value 47
Tabel 4.3.1 Tabel Nilai Balanced Accuracy 50
Tabel 4.3.2 Tabel Nilai G-Mean 52
Tabel 4.3.3 Tabel Nilai MAUC 54
Tabel 4.3.4 Tabel Nilai Sensitivity 56
Tabel 4.3.5 Tabel Nilai Specificity 57

INTISARI

Permasalahan ketidakseimbangan kelas pada dataset (*imbalanced dataset*) merupakan kondisi dimana nilai dari kelas minoritas sangat jauh lebih kecil dengan kelas mayoritas atau sangat kurang memadai sehingga model lebih mengenali pola pada kelas mayoritas dibanding kelas minoritas. Permasalahan tersebut merupakan salah satu tantangan yang sangat penting dalam penelitian *machine learning*, sehingga telah dikembangkan beberapa metode untuk mengatasinya. Namun metode-metode tersebut mayoritas hanya terfokus pada *binary dataset*, sehingga masih belum banyak metode yang terfokus pada *multiclass dataset*.

Penanganan *multiclass* tentu lebih sulit daripada *binary* karena melibatkan kelas yang lebih banyak. Untuk itu diperlukan algoritma yang memiliki fitur yang dapat mendukung penyesuaian terhadap kesulitan-kesulitan yang muncul pada *multiclass imbalanced dataset*, salah satu algoritma *ensemble* yang memiliki fitur-fitur untuk penyesuaian adalah algoritma XGBoost.

Pada 8 dari 9 dataset dengan metrik evaluasi *balanced accuracy, g-mean,* MAUC*, sensitivity, dan specificity*, algoritma XGBoost mampu mengungguli algoritma klasifikasi dan algoritma *ensemble* lain. Bahkan menyentuh nilai sempurna yaitu 1.00 pada dataset *new-thyroid*.

Kata Kunci: *Boosting*, XGBoost, Ketidakseimbangan Kelas, *Multiclass*, Klasifikasi.

ABSTRACT

Dataset imbalance problem is a condition where the number minority class is way smaller than the majority class or insufficient that the model recognizes patterns in the majority class more than the minority class. This problem is one of the most important challenges in machine learning research, so several methods have been developed to overcome it. However, the majority of these methods only focus on binary datasets, so there are still not many methods that focus on multiclass datasets.

Handling multiclass is certainly more difficult than binary because it involves more classes. For this reason, we need an algorithm that has features that can support adjustments to the difficulties that arise in multiclass imbalanced datasets, one of the ensemble algorithms that has features for adjustment is the XGBoost algorithm.

In 8 out of 9 datasets with evaluation metrics of balanced accuracy, g-mean, MAUC, sensitivity, and specificity, the XGBoost algorithm is able to outperform other classification algorithms and ensemble algorithms. It even hits a perfect score of 1.00 on the new-thyroid dataset.

***Keywords**: Boosting, XGBoost, Class Imbalance, Multiclass, Classification.*