TEKNIK BALANCING DATASET MENGGUNAKAN ALGORITMA ADASYN DAN BORDERLINE-SMOTE

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika

disusun oleh

TRIAS HANDAYANI

19.11.2733

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

TEKNIK BALANCING DATASET MENGGUNAKAN ALGORITMA ADASYN DAN BORDERLINE-SMOTE

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika

disusun oleh

TRIAS HANDAYANI

19.11.2733

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

HALAMAN PERSETUJUAN

SKRIPSI

TEKNIK BALANCING DATASET MENGGUNAKAN ALGORITMA ADASYN DAN BORDERLINE-SMOTE

yang disusun dan diajukan oleh

Trias Handayani

19.11.2733

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Februari 2023

Dosen Pembimbing,

Mardhiya Hayaty, S.T., M.Kom
NIK. 190302108

HALAMAN PENGESAHAN

SKRIPSI

TEKNIK BALANCING DATASET MENGGUNAKAN ALGORITMA ADASYN DAN BORDERLINE-SMOTE

yang disusun dan diajukan oleh

Trias Handayani

19.11.2733

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Februari 2023

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Norhikmah, M.Kom **NIK. 190302245**	
Erni Seniwati, S.Kom, M.Cs **NIK. 190302231**	
Mardhiya Hayaty, S.T., M.Kom. **NIK. 190302108**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Februari 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 : Trias Handayani
NIM : 19.11.2733

Menyatakan bahwa Skripsi dengan judul berikut:

Teknik Balancing Dataset Menggunakan Algoritma ADASYN dan Borderline-SMOTE

Dosen Pembimbing : Mardhiya Hayaty, S.T., M.Kom

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 Februari 2023

Yang Menyatakan,

Trias Handayani

HALAMAN PERSEMBAHAN

Puji syukur penulis ucapkan kehadirat Allah SWT yang melimpahkan karunia-Nya sehingga penulis mampu menyelesaikan penyusunan skripsi dengan judul "**Teknik Balancing Dataset Menggunakan Algoritma ADASYN dan Borderline-SMOTE**". Skripsi ini penulis persembahkan kepada:

1. Orang tua dan keluarga yang selalu memberikan semangat dan doa tanpa henti.
2. Ibu Mardhiya Hayaty, S.T., M.Kom selaku dosen pembimbing yang telah membantu dalam penyusunan skripsi.
3. Teman-teman kelas 19-IF-03 yang telah menemani dari awal hingga akhir semester yang telah membantu saya dalam menyelesaikan studi.
4. Teman-teman yang selalu mendengarkan keluh kesah dan memberikan motivasi untuk menyelesaikan skripsi.

KATA PENGANTAR

Puji syukur penulis ucapkan kehadirat Allah SWT yang melimpahkan karunia-Nya sehingga penulis mampu menyelesaikan penyusunan skripsi dengan judul "**Teknik Balancing Dataset Menggunakan Algoritma ADASYN dan Borderline-SMOTE**" sebagai syarat untuk menyelesaikan studi sarjana di Universitas Amikom Yogyakarta.

Dalam proses penyusunan skripsi ini penulis banyak mengalami kendala. Namun, berkat bantuan dari berbagai pihak skripsi ini dapat diselesaikan. Oleh karena itu, Penulis mengucapkan terimakasih kepada :

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku rektor Universitas Amikom Yogyakarta.
2. Ibu Mardhiya Hayaty, S.T., M.Kom selaku dosen pembimbing yang telah membimbing dalam penyusunan skripsi.
3. Orang tua dan keluarga yang telah mendoakan dan memberikan motivasi.
4. Seluruh dosen penguji yang telah memberikan saran sehingga sekripsi ini menjadi lebih baik.

Penulis berharap, skripsi ini dapat memberikan manfaat bagi para pembaca. Penulis menyadari bahwa skripsi ini terdapat banyak kekurangan sehingga saran dan kritik dibutuhkan untuk menyempurnakan skripsi ini.

Yogyakarta, 21 Februari 2023

Penulis

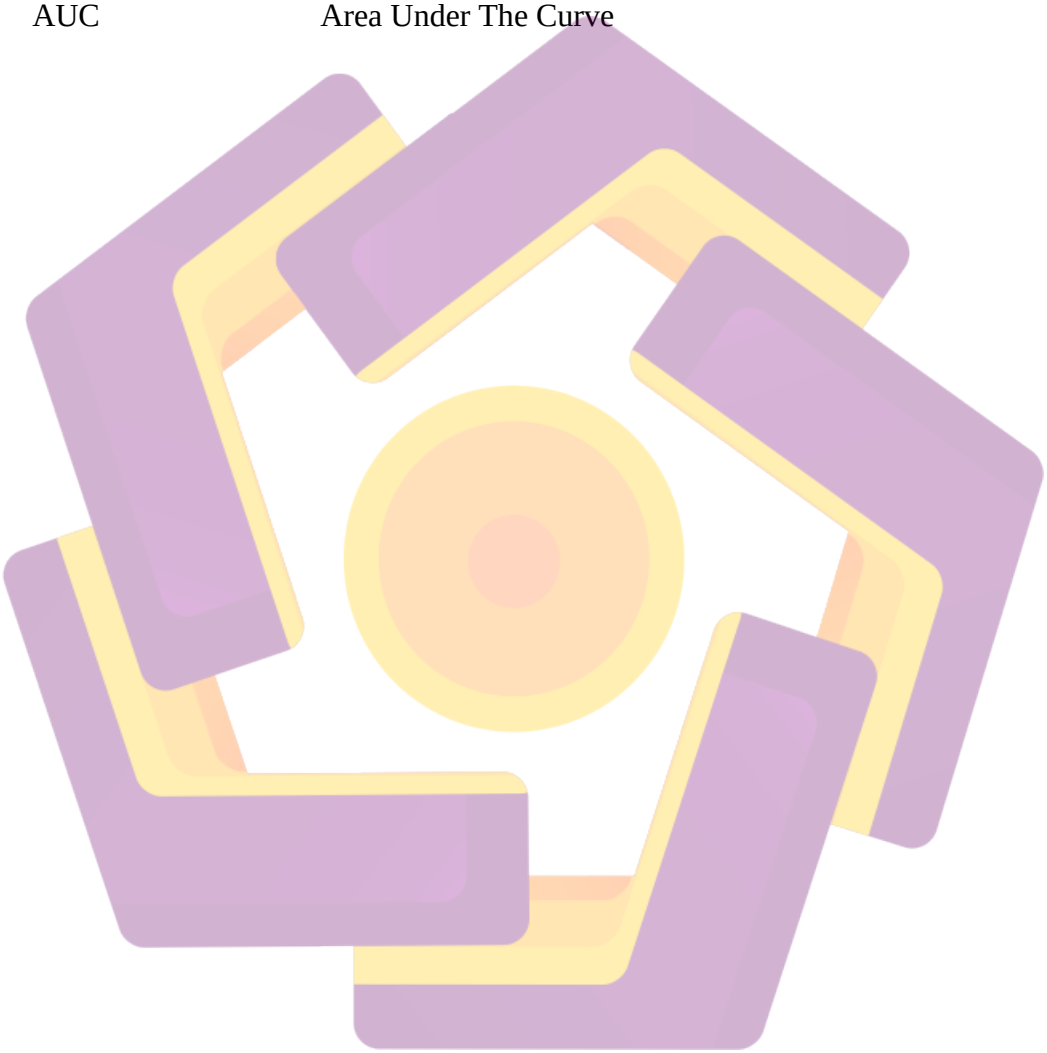
DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

ADASYN	Adaptive Synthetic
IR	Imbalance Ratio
ROC Curve	Receiver Operating Characteristic Curve
AUC	Area Under The Curve



INTISARI

Ketidakseimbangan dataset merupakan permasalahan umum yang sering ditemukan pada klasifikasi. Ketidakseimbangan dataset merupakan kondisi dimana sebuah dataset memiliki perbedaan jumlah data antara kelas satu dengan lainnya.

Masalah ketidakseimbangan ini mengakibatkan salah klasifikasi yang dapat menurunkan kinerja dari model klasifikasi. Penurunan kinerja model terjadi karena data lebih mengenali data kelas mayoritas yang lebih dominan sehingga kelas minoritas sering salah diklasifikasikan. Untuk mengatasi permasalahan klasifikasi ini, maka dataset diseimbangkan dengan meningkatkan data di kelas minoritas. Pada penelitian ini, ADASYN dan Borderline-SMOTE digunakan untuk menyeimbangkan dataset.

Dataset yang digunakan adalah dataset lung cancer yang tidak seimbang. Random Forest diterapkan sebagai algoritma klasikasi dengan nilai n_estimators=100 dan divalidasi menggunakan 5 stratified k-fold cross validation. Hasil penelitian menunjukan adanya peningkatan kinerja berdasarkan nilai AUC setelah dilakukan penyeimbangan baik menggunakan ADASYN maupun Borderline-SMOTE. Peningkatan kinerja berdasarkan nilai AUC yang terjadi antara 1%-13%. Pada beberapa k-fold ADASYN dan Borderline-SMOTE menghasilkan nilai AUC sebesar 100% yang menandakan bahwa model mampu mengklasifikasikan seluruh data dengan sempurna. Dari k-fold 2,4, dan 5 hasil nilai AUC menunjukan bahwa ADASYN lebih unggul dari Borderline-SMOTE.

Kata kunci: Data Tidak Seimbang, Oversampling, ADASYN, Borderline SMOTE, Random Forest.

ABSTRACT

Dataset imbalance is a common problem that is often found in classification. Dataset imbalance is a condition where a dataset has a difference in the amount of data between one class and another.

This imbalance problem results in misclassification which can reduce the performance of the classification model. The decline in model performance occurs because the data better recognizes the dominant majority class data so that the minority class is often misclassified. To overcome this classification problem, the dataset is balanced by increasing the data in the minority class. In this study, ADASYN and Borderline-SMOTE are used to balance the dataset.

The dataset used is an unbalanced lung cancer dataset. Random Forest is applied as a classicization algorithm with a value of n_estimators=100 and validated using 5 stratified k-fold cross validation. The results showed that there was an increase in performance based on the AUC value after balancing using both ADASYN and Borderline-SMOTE. Performance improvement based on the AUC value that occurs between 1% -13%. In several k-folds ADASYN and Borderline-SMOTE produce an AUC value of 100% which indicates that the model is able to classify all data perfectly. From k-fold 2, 4, and 5 the results of the AUC value show that ADASYN is superior to Borderline-SMOTE.

Keyword: Imbalance Data, Oversampling, ADASYN, Borderline SMOTE, Random Forest.