ANALISIS KINERJA ALGORITMA KLASIFIKASI BERBASIS POHON MENGGUNAKAN SMOTE TERHADAP DATA TIDAK SEIMBANG

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh
SIGIT WIBOWO
NIM. 1811016310020

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JULI 2023

ANALISIS KINERJA ALGORITMA KLASIFIKASI BERBASIS POHON MENGGUNAKAN SMOTE TERHADAP DATA TIDAK SEIMBANG

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh
SIGIT WIBOWO
NIM. 1811016310020

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JULI 2023

SKRIPSI

ANALISIS KINERJA ALGORITMA KLASIFIKASI BERBASIS POHON MENGGUNAKAN SMOTE TERHADAP DATA TIDAK SEIMBANG

Oleh :

SIGIT WIBOWO

1811015310020

Telah dipertahankan di depan Dosen Penguji pada tanggal 14 Juli 2023

Susunan Penguji :

Pembimbing I	**Dosen Penguji I**
Muhammad Itqan Mazdadi, S.Kom., M.Kom.	Dwi Kartini, S.Kom., M.Kom
NIP. 199006122019031013	NIP. 198704212012122003
Pembimbing II	**Dosen Penguji II**
Fatma Indriani, S.T., M.I.T., Ph.D	Friska Abadi, S.Kom., M.Kom
NIP. 198404202008122004	NIP. 198809132016121100 1

Banjarbaru, 22 Agustus 2023

Koordinator PS Ilmu Komputer,

Irwan Budiman, S.T., M.Kom

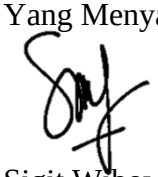
NIP. 197703252008121001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 23 Agustus 2023

Yang Menyatakan,



Sigit Wibowo

NIM. 1811016310020

ABSTRAK

ANALISIS KINERJA ALGORITMA KLASIFIKASI BERBASIS POHON MENGGUNAKAN SMOTE TERHADAP DATA TIDAK SEIMBANG

(Oleh: Sigit Wibowo; Pembimbing: Muhammad Itqan Mazdadi, S.Kom., M.Kom, dan Fatma Indriani, S.T.,M.I.T., Ph.D; 2023; 82 halaman).

Data tidak seimbang menjadi permasalahan tersendiri dalam proses penambangan data. Data tidak seimbang disebabkan distribusi kelas pada *dataset* tidak merata, sehingga terbentuk kelompok kelas mayoritas dan minoritas, dan berdampak pada hasil klasifikasi. SMOTE data menjadi upaya untuk meningkatkan kinerja model klasifikasi. Penerapan SMOTE pada algoritma berbasis pohon seperti *Random Forest,* CART, dan C5.0 dapat memperbaiki kinerja model dalam menghadapi *dataset* yang tidak seimbang dengan menambahkan sampel sintetis pada kelas minoritas, dan meningkatkan kemampuan klasifikasi pada kelas minoritas. Namun, diantara *Random Forest,* CART, dan C5.0 belum ada dilakukan perbandingan untuk menghadapi kasus data tidak seimbang. Pada penelitian ini melakukan analisis kinerja algoritma *Random Forest,* CART, dan C5.0 menggunakan SMOTE pada *dataset* Diabetes *Disesase*, Stroke *Prediction*, dan Customer Churn. Tujuan penelitian ini adalah mendapatkan pengetahuan baru mengenai model klasifikasi berbasis pohon yang dapat memberikan kinerja lebih baik pada data tidak seimbang berdasarkan perbandingan yang dilakukan. Setiap kinerja model dievaluasi menggunakan *recall*. Dalam analisis kinerja model, model *Random forest* dan CART tanpa penerapan SMOTE memiliki kinerja sama – sama baik, dengan *recall* rata-rata 48%, namun dalam penerapan SMOTE, *Random forest* dan C5.0 memiliki kinerja sama – sama baik dengan *recall* rata-rata 81%, CART. Penelitian ini menunjukan, SMOTE dapat meningkatkan nilai *recall* dari model yang digunakan. Namun pada hasil T-Test, CART, C5.0 dan *Random forest* tidak ada perbedaan signifikan antara model tanpa penerapan SMOTE dan model dengan penerapan SMOTE.

Kata kunci: C5.0, CART, Klasifikasi, *Random Forest,* SMOTE

ABSTRACT

PERFORMANCE ANALYSIS OF TREE-BASED CLASSIFICATION ALGORITHM USING SMOTE ON UNBALANCED DATA

(By: Sigit Wibowo; Advisor: Muhammad Itqan Mazdadi, S.Kom., M.Kom, and Fatma Indriani, S.T.,M.I.T., Ph.D; 2023; 82 pages).

Imbalanced data becomes a distinct problem in the data mining process. Imbalanced data is caused by an uneven class distribution in the dataset, resulting in majority and minority class groups, which can impact classification outcomes. SMOTE (Synthetic Minority Over-sampling Technique) is an effort to enhance the performance of classification models. Applying SMOTE to tree-based algorithms such as Random Forest, CART, and C5.0 can improve the model's performance when dealing with imbalanced datasets by adding synthetic samples to the minority class, thereby enhancing the classification ability of the minority class. However, no comparison has been made between Random Forest, CART, and C5.0 in handling imbalanced data cases. In this study, the performance of the Random Forest, CART, and C5.0 algorithms using SMOTE is analyzed on the Diabetes Disease, Stroke Prediction, and Customer Churn datasets. The research aims to gain new insights into tree-based classification models that can provide better performance on imbalanced data based on the conducted comparisons. The performance of each model is evaluated using recall. In performance analysis of the models, both Random Forest and CART models without the implementation of SMOTE exhibit similar good performance, with an average recall of 48%. However, in the application of SMOTE, both Random Forest and C5.0 models perform equally well with an average recall of 81%, surpassing CART. This study demonstrates that SMOTE can enhance the recall value of the utilized models.. Nevertheless, the results of the T-Test indicate that there is no significant difference between the models without the implementation of SMOTE and the models with the implementation of SMOTE for CART, C5.0, and Random Forest.

***Keywords**:* C5.0, CART, *Classification*, *Random Forest*, SMOTE

PRAKATA

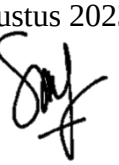
Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul "Analisis Kinerja Algoritma Kalsifikasi Berbasis Pohon Menggunakan SMOTE Terhadap Data Tidak Seimbang" guna memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

Penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang sangat mendukung penulis dapam penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut :

1. Keluarga yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini.
2. Bapak Muhammad Itqan Mazdadi, S.Kom., M.Kom., selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Ibu Fatma Indriani, S.T., M.I.T., Ph.D, selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Irwan Budiman S.T., M.Kom selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.

Akhir kata penulis menyadari bahwa penulisan ini jauh dari sempurna, namun penulis berharap bantuan berupa saran dan kritik yang membangun dari semua pihak guna perbaikan penulisan skripsi ini.

Banjarbaru, Agustus 2023



Sigit Wibowo

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN ii
PERNYATAAN iii
ABSTRAK iv
ABSTRACT v
PRAKATA vi
DAFTAR ISI vii
DAFTAR TABEL x
DAFTAR GAMBAR xiv
DAFTAR LAMPIRAN xv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan Penelitian 3
1.5 Manfaat Penelitian 3
BAB II TINJAUAN PUSTAKA 4
2.1 Kajian Terdahulu 4
2.2 Landasan Teori 8
2.2.1 Data Tidak Seimbang 8
2.2.2 *Synthetic Minority Oversampling Technique (*SMOTE*)* 8
2.2.3 C5.0 9
2.2.4 *Classification and Regression Tree (*CART*)* 11

2.2.5 *Random Forest* 13
2.2.6 *Hyperparameter Tuning* 14
2.2.7 Confusion Matrix 15
2.2.8 T-Test 17
BAB III METODE PENELITIAN 18
3.1 Alat Penelitian 18
3.2 Data 18
3.3 Prosedur Penelitian 18
3.3.1 Pengambilan Data 19
3.3.2 Preprocessing 19
3.3.3 Spliting Data 20
3.3.4 Resampling Data 20
3.3.5 Pembuatan Model 20
3.3.6 Evaluasi 20
3.3.7 Uji Signifikasi (T-Test) 20
BAB IV HASIL DAN PEMBAHASAN 21
4.1 Hasil 21
4.1.1 Pengumpulan *Dataset* 21
4.1.2 *Preprocessing* 23
4.1.3 Pembagian data 28
4.1.4 SMOTE 29
4.1.5 *Hyperparameter Tuning* 37
4.1.6 Klasifikasi dan Evaluasi Model 48
4.1.7 Uji Signifikasi (T-Test) 69
4.2 Pembahasan 72

BAB V PENUTUP..... 78

5.1 Kesimpulan 78

5.2 Saran..... 78

DAFTAR PUSTAKA

LAMPIRAN

DAFTAR TABEL

Tabel **Halaman**

Tabel 1 Keaslian penelitian 6
Tabel 2 Rancangan penelitian 7
Tabel 3 Confusion matrix 15
Tabel 4 Panduan penilaian AUC 16
Tabel 5 Hasil pengumpulan dataset 19
Tabel 6 Daftar atribut dataset Diabetes Disease 21
Tabel 7 Daftar atribut dataset Stroke Prediction 21
Tabel 8 Daftar atribut dataset Customer Churn 22
Tabel 9. Jumlah atribut masing - masing dataset 22
Tabel 10 Distribusi kelas dataset 23
Tabel 11 Hasil dari label encoding 24
Tabel 12 Dataset Diabetes Disease sebelum normalisasi 26
Tabel 13 Dataset Diabetes Disease sesudah normalisasi 26
Tabel 14 Dataset stroke sebelum normalisasi 27
Tabel 15 Dataset stroke sesudah normalisasi 27
Tabel 16 Dataset Customer Churn sebelum normalisasi 27
Tabel 17 Dataset Customer Churn sesudah normalisasi 28
Tabel 18 Pembagian data pada dataset Diabetes 28
Tabel 19 Pembagian data pada dataset Stroke 29
Tabel 20 Pembagian data pada dataset Customer Churn 29
Tabel 21 Distribusi kelas pada data latih 29
Tabel 22 Data sintetis baru pada data latih diabetes 31
Tabel 23 Data hasil SMOTE pada data latih diabetes 32
Tabel 24 Data sintesis baru pada data latih stroke 34
Tabel 25 Data hasil SMOTE pada data latih stroke 34
Tabel 26 Data sintesis baru pada data latih Customer Churn 36
Tabel 27 Data hasil SMOTE pada data latih Customer Churn 36
Tabel 28 Parameter yang digunakan pada hyperparameter tuning 37

Tabel 29 Pencarian hyperparameter model dengan grid search..... 38
Tabel 30 Penentuan nilai parameter optimal RF pada dataset diabetes..... 39
Tabel 31 Penentuan nilai parameter optimal RF+SMOTE pada dataset diabetes 39
Tabel 32 Penentuan nilai parameter optimal RF pada dataset stroke 40
Tabel 33 Penentuan nilai parameter optimal RF+SMOTE pada dataset stroke 40
Tabel 34 Penentuan nilai parameter optimal RF pada dataset Customer Churn..... 41
Tabel 35 nilai parameter optimal RF+SMOTE pada dataset Customer Churn 41
Tabel 36 Penentuan nilai parameter optimal CART pada dataset diabetes 42
Tabel 37 Penentuan parameter optimal CART+SMOTE pada dataset diabetes..... 42
Tabel 38 Penentuan nilai parameter optimal CART pada dataset stroke..... 43
Tabel 39 Penentuan nilai parameter optimal CART+SMOTE pada dataset stroke.. 43
Tabel 40 Penentuan nilai parameter optimal CART pada dataset Customer Churn.. 44
Tabel 41 Penentuan parameter optimal CART+SMOTE pada dataset Customer 44
Tabel 42 Penentuan nilai parameter optimal C5.0 pada dataset diabetes 45
Tabel 43 Penentuan nilai parameter optimal C5.0+SMOTE pada dataset diabetes .. 45
Tabel 44 Penentuan nilai parameter optimal C5.0 pada dataset stroke..... 46
Tabel 45 Penentuan nilai parameter optimal C5.0+SMOTE pada dataset stroke..... 46
Tabel 46 Penentuan nilai parameter optimal C5.0 pada dataset Customer Churn..... 47
Tabel 47 Penentuan nilai parameter optimal C5.0+SMOTE pada dataset..... 47
Tabel 48 Contoh data set Diabetes Disease 49
Tabel 49 Hasil perhitungan perhitungan entropy, information gain, dan gain ratio.. 50
Tabel 50. Confusion matrix C5.0 tanpa SMOTE..... 51
Tabel 51 Confusion matrix C5.0 dengan SMOTE..... 51
Tabel 52 Kemungkinan pemilah pada variabel Pregnancies 52
Tabel 53 Confusion matrix CART tanpa SMOTE..... 52
Tabel 54 Confusion Matrix CART dengan SMOTE 53
Tabel 55 Dataset diabetes untuk random forest 53
Tabel 56 Nilai entropy pada pohon pertama 54
Tabel 57 Confusion matrix random forest 55
Tabel 58 Confusion matrix random forest dengan SMOTE 55
Tabel 59 Contoh tabel dataset stroke 56

Tabel 60 Hasil perhitungan perhitungan entropy, information gain, dan gain ratio .. 57
Tabel 61 Confusion matrix C5.0 tanpa SMOTE 58
Tabel 62 Confusion matrix C5.0 dengan SMOTE 58
Tabel 63 Kemungkinan pemilah pada variabel age 59
Tabel 64 Confusion matrix CART tanpa SMOTE 59
Tabel 65 Confusion matrix CART dengan SMOTE 60
Tabel 66 Contoh tabel dataset stroke 60
Tabel 67 nilai *entropy* dan *gain* pada *root node* 61
Tabel 68 Confusion matrix Random forest 62
Tabel 69 Random forest dengan SMOTE 62
Tabel 70 Contoh dataset Customer Churn 63
Tabel 71 hasil perhitungan entropy, information gain, dan gain ratio 64
Tabel 72 Confusion martix C5.0 65
Tabel 73 Confusion martix C5.0 dengan SMOTE 65
Tabel 74 Kemungkinan pemilah pada variabel gender 66
Tabel 75 Confusion matrix CART 66
Tabel 76 Confusion matrix CART dengan resampling 67
Tabel 77 Contoh dataset Customer Churn 67
Tabel 78 nilai entropy dan gain pada root node 68
Tabel 79 Confusion matrik Random forest 69
Tabel 80 Confusion matrik Random forest dengan resampling 69
Tabel 81 Hasil T-Test model random forest 70
Tabel 82 Hasil T-Test model CART 70
Tabel 83 Hasil T-Test model C5.0 70
Tabel 84 Uji T pada model Rf dan CART tanpa SMOTE 71
Tabel 85 Uji T pada model Random forest dan C5.0 tanpa penerapan SMOTE 71
Tabel 86 Uji T pada model CART dan C5.0 tanpa penerapan SMOTE 71
Tabel 87 Uji T pada model Rf dan CART dengan penerapan SMOTE 72
Tabel 88 Uji T pada model Random forest dan C5.0 dengan penerapan SMOTE 72
Tabel 89 Uji T pada model CART dan C5.0 dengan penerapan SMOTE 72
Tabel 90 Recall klasifikasi Diabetes Disease, Stroke Prediction, dan IT Customer

Churn 74
Tabel 91 Hasil uji signifikasi 77

DAFTAR GAMBAR

Gambar **Halaman**

Gambar 1 Cara kerja SMOTE (Aldraimli et al., 2020) 8
Gambar 2 *Flowchart* SMOTE (Sopiyan et al., 2022) 9
Gambar 3 *Flowchart* C5.0 10
Gambar 4 *Flowchart* CART 12
Gambar 5 ilustrasi Random Forest 13
Gambar 6 *Flowchart* random forest 14
Gambar 7 Alur penelitian 19
Gambar 8 Distribusi kelas setiap *dataset* 23
Gambar 9 Visualisasi data hasil SMOTE pada data latih diabetes 31
Gambar 10 Visualisasi data hasil SMOTE pada data latih stroke 34
Gambar 11 Visualisasi data hasil SMOTE pada data latih Customer Churn 36
Gambar 12 Diagram alur model 48
Gambar 13 Contoh hasil pembentukan cabang di *node* 1 50
Gambar 14. *Node* 1 random forest 54
Gambar 15 Perbandingan nilai *recall* pada *dataset* diabetes 55
Gambar 16 Contoh pembentukan cabang *node* 1 57
Gambar 17 Contoh pembentukan cabang *node* 1 61
Gambar 18 Perbandingan kinerja model pada *dataset* Stroke 62
Gambar 19 Contoh pembentukan *node* 1 64
Gambar 20 *Node* 1 pada *dataset* Customer Churn 68
Gambar 21 Perbandingan kinerja model pada *dataset* Customer Churn 69
Gambar 22 Rata - rata *recall* model 75

DAFTAR LAMPIRAN

Lampiran 1. Layout implementasi model

Lampiran 2. Layout cross validation

Lampiran 3. Layout proses cross validation

Lampiran 4. Konfigurasi parameter