

SKRIPSI



**IMPLEMENTASI ALGORITMA *RANDOM FOREST* DALAM
PREDIKSI *STROKE* DENGAN PENGGUNAAN *SYNTHETIC
MINORITY OVER-SAMPLING TECHNIQUE***

JOHANES GERALD

NIM. 2010511106

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAKARTA
2023**

SKRIPSI



**IMPLEMENTASI ALGORITMA *RANDOM FOREST* DALAM
PREDIKSI *STROKE* DENGAN PENGGUNAAN *SYNTHETIC
MINORITY OVER-SAMPLING TECHNIQUE***

JOHANES GERALD

NIM. 2010511106

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAKARTA**

2023

PERNYATAAN ORISINALITAS

Skripsi ini merupakan hasil karya sendiri serta semua sumber referensi yang dikutip maupun yang dirujuk telah saya nyatakan benar.

Nama : Johanes Gerald
NIM : 2010511106
Tanggal : 29 November 2023

Bilamina di kemudian hari ditemukan ketidaksesuaian dengan pernyataan ini, maka saya bersedia dituntut dan diproses sesuai dengan ketentuan berlaku.

Jakarta, 30 November 2023

Yang Menyatakan,



Johanes Gerald

**PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK
KEPENTINGAN AKADEMIS**

Sebagai civitas akademik Universitas Pembangunan Nasional Veteran Jakarta, saya yang bertanda tangan di bawah ini:

Nama : Johanes Gerald
NIM : 2010511106
Fakultas : Ilmu Komputer
Program Studi : S1 Informatika

Demi pembangunan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Pembangunan Nasional Veteran Jakarta Hak Bebas Royalti Eksklusif (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul:

**IMPLEMENTASI ALGORITMA *RANDOM FOREST* DALAM PREDIKSI
STROKE DENGAN PENGGUNAAN *SYNTHETIC MINORITY OVER-
SAMPLING TECHNIQUE***

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti ini Universitas Pembangunan Nasional Veteran Jakarta berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan Skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sesungguhnya.

Dibuat di : Jakarta
Pada Tanggal : 30 November
2023

Yang menyatakan,



Johanes Gerald

LEMBAR PENGESAHAN


Tugas akhir ini diajukan oleh:

Nama : Johanes Gerald
NIM : 2010511106
Program Studi : S1-Informatika
Judul : IMPLEMENTASI ALGORITMA *RANDOM FOREST*
DALAM PREDIKSI *STROKE* DENGAN
PENGUNAAN *SYNTHETIC MINORITY OVER-*
SAMPLING TECHNIQUE

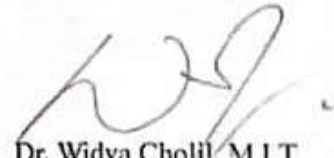
Telah berhasil dipertahankan dihadapan Tim Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana pada Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta.




Dr. Ermatita, M.Kom
Penguji 1




Ika Nurfauli Isnainiyah, S.Kom., M.Sc.
Penguji 2



Dr. Widya Choiil, M.I.T
Pembimbing



Prof. Dr. Ir. Supriyanto, ST., M.Sc., IPM
Dekan



Dr. Widya Choiil, M.I.T
Kepala Program Studi

Ditetapkan di : Jakarta
Tanggal Ujian : 10 Januari 2024

IMPLEMENTASI ALGORITMA *RANDOM FOREST* DALAM PREDIKSI *STROKE* DENGAN PENGGUNAAN *SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE*

Johanes Gerald

ABSTRAK

Stroke adalah salah satu masalah kesehatan global yang signifikan dan menjadi penyebab utama kecacatan serta kematian di seluruh dunia. Di Indonesia, penyakit *stroke* telah menjadi salah satu penyumbang penyakit paling mematikan. Menurut data pada profil kesehatan Indonesia tahun 2020, *stroke* menempati posisi ketiga dengan jumlah kasus sebanyak 1.789.261. Tujuan penelitian ini adalah untuk mengidentifikasi pasien yang berisiko tinggi terkena *stroke*. Algoritma yang digunakan adalah *Random Forest Classifier* dengan menggunakan *Synthetic Minority Oversampling Technique* untuk menyeimbangkan kelas data. Pada metode *Random Forest* menggunakan *Synthetic Minority Oversampling Technique* didapatkan hasil akurasi sebesar 95,61%, nilai presisi sebesar 93,66%, nilai recall sebesar 97,85%, dan nilai f1-score sebesar 95,71%. Sedangkan untuk model *Random Forest* didapatkan hasil akurasi sebesar 90,15%, nilai presisi sebesar 90,5%, nilai recall sebesar 90,15%, dan nilai f1-score sebesar 90,32%. Karena terjadi *imbalance class* maka algoritma *Random Forest* saja tidak tepat digunakan tanpa melakukan *resampling*. Sehingga *Random Forest – SMOTE* dapat digunakan sebagai salah satu algoritma untuk memprediksi *stroke*.

Kata kunci: Klasifikasi, *Stroke*, *Random Forest*, SMOTE.

**IMPLEMENTATION OF RANDOM FOREST ALGORITHM IN STROKE
PREDICTION USING SYNTHETIC MINORITY OVER-SAMPLING
TECHNIQUE**

Johanes Gerald

ABSTRACT

Stroke is one of the significant global health issues, being a leading cause of disability and death worldwide. In Indonesia, stroke has emerged as one of the most fatal diseases. According to the 2020 Indonesia health profile data, stroke ranks third with a total of 1,789,261 reported cases. The aim of this research is to identify patients at high risk of stroke. The algorithm employed is the Random Forest Classifier utilizing the Synthetic Minority Oversampling Technique to balance the class data. In the Random Forest method using the Synthetic Minority Oversampling Technique, the results showed an accuracy of 95.61%, precision of 93.66%, recall of 97.85%, and an f1-score of 95.71%. Meanwhile, for the Random Forest model, the accuracy was 90.15%, precision was 90.5%, recall was 90.15%, and the f1-score was 90.32%. Due to class imbalance, using the Random Forest algorithm alone is not suitable without resampling. Therefore, Random Forest – SMOTE can be utilized as one of the algorithms to predict strokes.

Keywords: Classification, Stroke, Random Forest, SMOTE.

KATA PENGANTAR

Puji Syukur penulis panjatkan kehadiran Tuhan Yesus Kristus, yang senantiasa memberikan petunjuk, kelancaran serta kemudahan kepada penulis dalam menyelesaikan skripsi ini, serta kasih anugerahNya yang tidak pernah berkesudahan. Penulisan skripsi ini bertujuan untuk memenuhi salah satu prasyarat untuk memperoleh gelar Sarjana Komputer, Jurusan Informatika.

Dalam penulisan skripsi ini, penulis mendapat banyak dukungan serta bantuan dari berbagai pihak, baik berupa materi, spiritual, dan informasi. Pada kesempatan kali ini, penulis mengucapkan terimakasih kepada:

1. Tuhan Yesus Kristus.
2. Kedua orang tua, cici serta keluarga tercinta yang selalu mendoakan serta mendukung penulis sehingga dapat menyelesaikan skripsi ini.
3. Ibu Dr. Widya Cholil, M.I.T, selaku dosen pembimbing dan Kaprodi Informatika yang berjasa dan memberikan bimbingan hingga terselesaikannya skripsi ini
4. Bapak Prof. Dr. Ir. Supriyanto, ST., M.Sc., IPM selaku Dekan Fakultas Ilmu Komputer.
5. Ibu Dr. Ermatita, M.Kom dan Ibu Ika Nurlaili Isnainiyah, S.Kom., M.Sc selaku dosen penguji.
6. Seluruh jajaran Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta yang telah membantu dalam perizinan dan administrasi.
7. Teman – teman Informatika yang berjuang bersama selama perkuliahan, memberikan semangat dan dorongan untuk dapat menyelesaikan kuliah dan skripsi ini.
8. Kakak tingkat penulis yang seringkali menjadi tempat berdiskusi yaitu Bang Ojan, Bang Kipau, Bang Radit.
9. Prambanan Crew yaitu Gilbert, Tito, Sarah, Nida, Billy, Wildan dan sahabat – sahabat perjuangan yang selalu membantu dan memberikan masukan serta doa kepada penulis agar terselesaikannya skripsi ini dengan baik.
10. Terima kasih banyak kepada semua pihak atas bantuan, dukungan, semangat, dan doa yang tidak dapat penulis sebutkan satu persatu

11. Johannes Gerald, *last but not least*, ya! diri saya sendiri. Apresiasi sebesar-besarnya karena telah bertanggung jawab untuk menyelesaikan apa yang telah dimulai. Terima kasih karena terus berusaha dan tidak menyerah, serta senantiasa menikmati setiap prosesnya yang bisa dibilang tidak mudah. Terima kasih sudah bertahan.

Penyusun menyadari bahwa penyusunan skripsi ini masih jauh dari sempurna, oleh karena itu penyusun mengharapkan kritik dan saran yang bersifat membangun untuk kesempurnaan skripsi ini. Akhir kata penyusun mengharapkan semoga skripsi ini dapat bermanfaat bagi semua pihak.

Jakarta, 10 Januari 2024

Penulis

DAFTAR ISI

LEMBAR JUDUL.....	i
PERNYATAAN ORISINALITAS	ii
PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS.....	iii
LEMBAR PENGESAHAN.....	iv
ABSTRAK	v
<i>ABSTRACT</i>	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xiv
DAFTAR LAMPIRAN	xv
BAB I PENDAHULUAN	1
I.1 Latar Belakang	1
I.2 Rumusan Masalah	3
I.3 Tujuan Penelitian	3
I.4 Manfaat Penelitian.....	3
I.5 Batasan Masalah.....	4
BAB II TINJAUAN PUSTAKA.....	5
II.1 <i>Stroke</i>	5
II.2 <i>Data Mining</i>	5
II.3 Klasifikasi	7
II.4 Algoritma <i>Random Forest</i>	8
II.5 <i>Imbalance Data</i>	11
II.6 <i>Oversampling</i>	11
II.7 SMOTE.....	12
II.8 <i>Python</i>	13
II.9 Evaluasi.....	13
II.9.1 <i>Accuracy</i>	14
II.9.2 <i>Precision</i>	14

II.9.3 <i>Recall</i>	14
II.9.4 <i>Specificity</i>	14
II.9.5 <i>F1-score</i>	15
II.10 Penelitian Relevan	15
BAB III METODOLOGI PENELITIAN.....	18
III.1 Kerangka Berpikir	18
III.1.1 Studi Literatur	19
III.1.2 Perumusan Masalah	19
III.1.3 Pengumpulan Data	19
III.1.4 <i>Preprocessing</i>	20
III.1.5 <i>Exploratory Data Analysis</i>	20
III.1.6 Pembagian Data	21
III.1.7 <i>Oversampling SMOTE</i>	21
III.1.8 Pembuatan Model	21
III.1.9 Evaluasi.....	22
III.1.10 <i>Testing</i>	22
III.2 Alat Pendukung Penelitian	23
III.2.1 Perangkat Keras	23
III.2.2 Perangkat Lunak	23
III.3 Tempat dan Waktu Penelitian.....	23
III.4 Jadwal Penelitian	23
BAB IV HASIL DAN PEMBAHASAN	24
IV.1 Data	24
IV.2 <i>Preprocessing</i>	25
IV.2.1 <i>Feature Selection</i>	25
IV.2.2 Categorical Encoding.....	25
IV.2.3 <i>Mean Imputation</i>	28
IV.3 <i>Exploratory Data Analysis</i>	29
IV.4 Pembagian Data.....	39
IV.5 <i>Oversampling SMOTE</i>	39
IV.6 Pembuatan Model.....	40
IV.7 Evaluasi	43

IV.8 <i>Testing</i>	47
BAB V PENUTUP	49
V.1 Kesimpulan	49
V.2 Saran	49
DAFTAR PUSTAKA	xvi
LAMPIRAN	xix

DAFTAR GAMBAR

Gambar II.1 Tahapan Klasifikasi	8
Gambar II.2 Alur kerja Algoritma Random Forest	10
Gambar II.3 Ilustrasi Algoritma SMOTE (Vijayvargiya et al., 2021)	12
Gambar III.1 Diagram Kerangka Berpikir	18
Gambar IV.1 Exploded Pie Chart data stroke dan non-stroke	25
Gambar IV.2 Source Code feature selection	25
Gambar IV.3 Source Code Categorical Encoding.....	26
Gambar IV.4 Tipe data sebelum categorical encoding	26
Gambar IV.5 Tipe data setelah categorical encoding.....	26
Gambar IV.6 Kolom mengandung Null Value.....	28
Gambar IV.7 Source Code Mean Imputation.....	28
Gambar IV.8 Kolom setelah Mean Imputation	29
Gambar IV.9 Heatmap korelasi antar variabel	29
Gambar IV.10 Grafik Feature Importance	30
Gambar IV.11 Hexagonal Bining distribusi data BMI.....	31
Gambar IV.12 Hexagonal Bining distribusi data age.....	31
Gambar IV.13 Hexagonal Bining distribusi data avg_glucose level	32
Gambar IV.14 Source Code Spliiting Data	39
Gambar IV.15 Source Code Oversampling SMOTE	40
Gambar IV.16 Install pustaka.....	40
Gambar IV.17 Source Code Penggabungan DataFrame	41
Gambar IV.18 Source Code Penyiapan Environment.....	41
Gambar IV.19 Source Code Pembuatan Model	41
Gambar IV.20 Source Code Cross-Validation	41
Gambar IV.21 Hasil Cross-Validation Random Forest.....	42
Gambar IV.22 Hasil Cross-Validation SMOTE-Random Forest.....	42
Gambar IV.23 Source Code Membuat Predict Model	42
Gambar IV.24 Pohon Model Random Forest - SMOTE.....	42
Gambar IV.25 Confusion Matrix	44
Gambar IV.26 ROC Curve	45

Gambar IV.27 Error Plot.....	46
Gambar IV.28 Learning Curve.....	47
Gambar IV.29 Prediksi Model “STROKE”	48
Gambar IV.30 Prediksi Model “NON-STROKE”	48

DAFTAR TABEL

Tabel II.1 Confusion Matrix	13
Tabel III.1 Informasi Atribut Dataset	19
Tabel III.2 Pertanyaan Kuesioner	22
Tabel III.3 Jadwal Penelitian	23
Tabel IV.1 Tabel Dataset Stroke	24
Tabel IV.2 Perubahan data setelah Categorical Encoding	27
Tabel IV.3 Dataset setelah Label Encoder	27
Tabel IV.4 Tabel data percobaan	32
Tabel IV.5 Tabel data “Ya” heart_disease	35
Tabel IV.6 Tabel data “No” heart_disease	37
Tabel IV.7 Tabel distribusi kelas stroke sebelum dan sesudah oversampling	40
Tabel IV.8 Perbandingan hasil evaluasi model	43

DAFTAR LAMPIRAN

Lampiran 1. Lembar Hasil Turnitin	xix
Lampiran 2. Source Code.....	xxvii
Lampiran 3. Kuesioner Data Testing	xxxii
Lampiran 4. Hasil Kuesioner Data Testing.....	xxxv