Jurnal TICOM: Technology of Information and Communication
Volume 13, Nomor 3, Mei 2025 Hal: 141-146 E-ISSN: 2962-7982

Deteksi Dini Penyakit Stroke pada Data Tidak Seimbang Menggunakan SMOTE dan Random Forest

Muhammad Iqbal Aryabima[1], Rusdah[2*], Ririt Roeswidiah[3], Ahmad Pudoli[4]

[1]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
[2] Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
[3,4]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Jl. Ciledug Raya, Petukangan Utara, Pesanggrahan, Jakarta Selatan
Email: [1]aryabal517@gmail.com, [2*]rusdah@budiluhur.ac.id, [3]ririt.roeswidiah@budiluhur.ac.id,
[4]ahmad.pudoli@budiluhur.ac.id
(* : coressponding author)

Abstrak—**Hilangnya sirkulasi darah ke bagian otak menyebabkan stroke, yang kemudian dikenal juga dengan serangan otak. Selain itu, pembekuan gumpalan darah juga merupakan penyebab utama stroke di otak. Stroke merupakan salah satu dari 10 penyakit paling mematikan di Indonesia. Menurut laporan hasil Riskesdas 2018, angka prevalensi stroke nasional masih cukup tinggi yaitu 10,9% per 1000 penduduk di Indonesia. Penelitian ini bertujuan untuk deteksi dini penyakit stroke dengan menerapkan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) dan menggunakan *Random Forest.* Data yang digunakan bersifat *public* dari website www.kaggle.com dengan total 4981 *record* yang terdiri dari 11 atribut. Komposisi data tidak seimbang dengan 4733 data negatif stroke (95%) dan 248 positif stroke (5%). Penanganan *imbalanced data* menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Hasil SMOTE membentuk komposisi data menjadi 79% data negative stroke dan 21% data positif stroke. Hasil penelitian menunjukkan bahwa SMOTE dapat meningkatkan performa model Random Forest.**

Kata Kunci—***Random Forest, Klasifikasi,* Penyakit Stroke, SMOTE*, Imbalanced Dataset***

Abstract— ***Loss of blood circulation to the brain causes stroke, which is also known as a brain attack. In addition, blood clots are also the main cause of stroke in the brain. Stroke is one of the 10 most deadly diseases in Indonesia. According to the 2018 Riskesdas report, the national stroke prevalence rate is still quite high, namely 10.9% per 1000 people in Indonesia. This study aims to detect stroke early by applying the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology and using Random Forest. The data used is publicly available from the website www.kaggle.com, comprising a total of 4981 records with 11 attributes. The data composition is unbalanced with 4733 negative stroke data (95%) and 248 positive stroke (5%). Handling imbalanced data using Synthetic Minority Oversampling Technique (SMOTE). The results of SMOTE form the data composition into 79% negative stroke data and 21% positive stroke data. The study's results showed that SMOTE can enhance the performance of the Random Forest model.***

Keywords: Random Forest, Classification, Stroke Disease, SMOTE, Imbalanced Data.

I. PENDAHULUAN

Penyakit stroke disebabkan oleh terputusnya sirkulasi darah ke bagian otak, yang kemudian dikenal juga dengan serangan otak. Selain itu, pembekuan gumpalan darah juga merupakan penyebab utama stroke di otak. Pembuluh darah yang mengantarkan bagian otak kemudian kekurangan darah dan oksigen. Sel-sel otak akan mati akibat kekurangan darah dan oksigen, dan bagian tubuh yang diaturnya akan berhenti bekerja [1].

Menurut laporan hasil Riskesdas 2018, angka prevalensi stroke nasional masih cukup tinggi yaitu 10,9% per 1000 penduduk di Indonesia. Provinsi di Indonesia dengan angka prevalensi stroke tertinggi adalah Kalimantan Timur yaitu sebesar 14,7% per 1000 penduduk, disusul provinsi Daerah Istimewa Yogyakarta yaitu sebesar 14.6% per 1000 penduduk. Penduduk yang berusia diatas 75 tahun menempati urutan pertama penderita stroke dengan angka prevelensi 50.2% per 1000 penderita [2].

Penelitian sebelumnya yang bertujuan untuk mendiagnosis stroke dilakukan dengan menggunakan Naïve Bayes [3], *K-Nearest Neighbor* [4], *Classification and Regression Tree (CART)* [5], *Decision Tree C.45* [6], dan *Random Forest* [1].

Pada penelitan ini menggunakan *dataset* yang bersifat *public*. Namun permasalahan yang sering muncul pada *dataset public* ialah data yang digunakan tidak seimbang cenderung ke arah positif ataupun ke arah negatif. Pada umumnya, ketika algoritma *machine learning* menerima *dataset* yang tidak seimbang, hal ini dapat mengakibatkan model memiliki tingkat sensitivitas yang rendah terhadap kelas minoritas. Dampaknya ialah model cenderung tidak mampu mengidentifikasi dengan akurat data dari kelas minoritas [7].

Synthetic Minority Oversampling Technique (SMOTE) ialah salah satu pendekatan teknik oversampling yang digunakan untuk menangani ketidakseimbangan data [8]. Dalam kasus *dataset* yang tidak seimbang, di mana jumlah sampel pada kelas minoritas lebih sedikit dibandingkan dengan kelas mayoritas, SMOTE efektif digunakan untuk mengatasi masalah ini. SMOTE bekerja dengan cara menghasilkan data

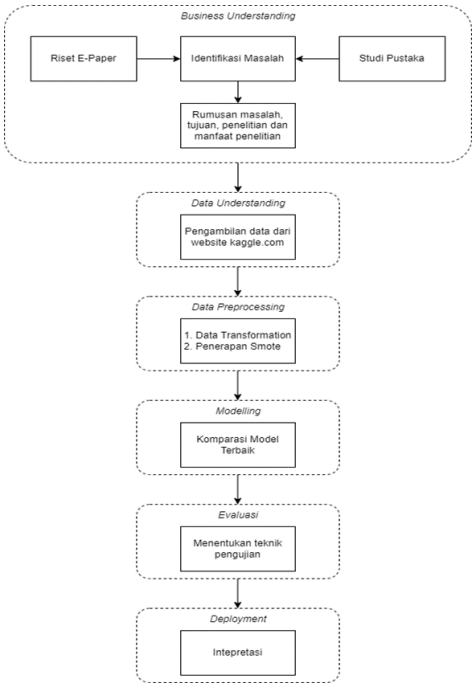
sintetis tambahan untuk kelas minoritas sehingga dapat seimbang dengan kelas mayoritas [7].

Data mining merupakan istilah yang digunakan untuk menggambarkan proses penemuan pengetahuan dalam sebuah *database*. Proses ini melibatkan penggunaan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengidentifikasi dan mengekstraksi informasi yang berharga serta pengetahuan dari berbagai *database* yang berskala besar [6].

Klasifikasi merupakan suatu tugas yang melibatkan pembelajaran atau pelatihan terhadap fungsi target yang memetakan setiap set atribut atau fitur ke salah satu label kelas yang tersedia yang dapat diartikan sebagai proses penilaian objek data untuk menempatkannya ke dalam kelas tertentu dari sekumpulan kelas yang ada [9].

Penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dan bertujuan untuk membuat model diagnosis stroke pada kasus data tidak seimbang.

II. METODE PENELITIAN



Gambar 1 Tahapan Penelitian

Gambar 1 menunjukkan tahapan penelitian yang menerapkan metodologi CRISP-DM. CRISP-DM adalah suatu kerangka kerja yang memungkinkan interpretasi masalah bisnis ke dalam teknik data mining. Dirancang untuk melaksanakan tugas data mining secara mandiri pada berbagai aplikasi dan teknologi, CRISP-DM mewakili proses standar lintas industri dalam domain data mining. Fokus utamanya adalah pada implementasi yang terfokus pada industri, dan prosesnya telah diadopsi secara luas sebagai bagian integral dari *Knowledge Discovery* [10]. Berikut adalah penjelasan dari setiap tahapan pada penelitian ini:

A. Pemahaman Bisnis *(Business Understanding)*

Pada tahap pertama dilakukan identifikasi masalah dengan cara studi pustaka. Beberapa penelitian terdahulu terkait diagnosis penyakit stroke menggunakan teknik data mining telah dipelajari untuk mengidentifikasi masalah dan menentukan tujuan penelitian. Masalah yang diidentifikasi adalah bagaimana membuat model prediksi penyakit stroke pada kasus dataset tidak seimbang.

B. Pemahaman Data *(Data Understanding)*

Setelah masalah dirumuskan, kemudian dilakukan pengumpulan data pada halaman *webite Kaggle.com* dengan judul *Brain Stroke Prediction Dataset*. *Dataset* tersebut bersifat *public* dengan jumlah *record* 4981 data yang terdiri dari 11 *attributes* dengan total 4733 data negatif stroke dan 248 positif stroke. Dengan demikian, komposisi dataset yang digunakan tidak imbang, yaitu 95% negatif stroke dan 5% positif stroke.

C. Persiapan Data *(Data Preparation)*

Data pada penelitian ini menggunakan data bersifat *public* yang bersumber dari website Kaggle.com yang sudah bersih tidak ada *missing value*. Pada tahap data *preprocessing* dilakukan transformasi yaitu normalisasi data pada beberapa atribut. Kemudian dilakukan *discretize* dengan data yang bertipe data nominal. Data yang bertipe nominal dipetakan ke dalam rentang nilai yang sudah ditentukan.

Dataset pada penelitian ini memiliki data tidak seimbang atau *data imbalanced* dan cenderung ke arah negatif stroke. Penanganan data tidak imbang menggunakan metode SMOTE oversampling.

D. *Modelling*

Pada tahap *modelling* dilakukan eksplorasi yang bertujuan untuk mencari model terbaik. Tahap validasi menggunakan metode *cross validation* dan *splitting data* serta dilakukannya juga komparasi algoritma diantaranya *Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) dan Random Forest (RF)*. Eksplorasi metode penentuan data training dan data testing ini dilakukan untuk mengetahui komposisi data terbaik, setelah teknik SMOTE diterapkan pada dataset.

Random Forest adalah metode klasifikasi yang terdiri dari kumpulan pohon keputusan (*decision tree*) terstruktur. Setiap pohon keputusan (*decision tree*) dalam *Random Forest* memberikan suara untuk kelas yang paling populer berdasarkan vektor acak yang didistribusikan secara identik [11]. Random Forest dikenal sebagai algoritma dengan kemampuan yang dapat menangani data tidak seimbang [11]. Berikut persamaan (1) dan (2) merupakan perhitungan algoritma *Random Forest.*

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2 pi \quad (1)$$

Keterangan :

- *Entropy*: S = Himpunan Kasus
- N = Jumlah Partisi S
- pi = Proporsi dari Si terhadap S

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

Keterangan :
- Gain: S = Himpunan Kasus.
- A = Atribut
- n = Jumlah Partisi Atribut A
- | S_i | = Jumlah Kasus pada partisi ke-i
- | S | = Jumlah Kasus dalam S

Untuk hasil akhir digunakan *gain ratio*, sedangkan untuk mendapatkan nilai *gain ratio*, dilakukan perhitungan *split information* sesuai dengan persamaan (3) dan (4).

$$Split\ Information = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2(\frac{|D_j|}{|D|}) \quad (3)$$

Keterangan :
|D*i*|/|D| = Probabilitas kelas *i*

$$Gain\ Ratio = \frac{Gain\ (S,A)}{Split\ Info(S,A)} \quad (4)$$

E. Evaluasi

Pada tahap ini dilakukan evaluasi menggunakan *confusion matrix* untuk dengan parameter uji akurasi, presisi, *recall* dan *area under the curve* (AUC). Confusion Matrix digunakan untuk menganalisis seberapa baik classifier mengenali data pada kelas yang berbeda [12]. Tabel I adalah ilustrasi dari *confusion matrix* [13].

TABEL I
TABEL *CONFUSION MATRIX*

Nilai Prediksi	**Nilai Aktual**	
	Postive	***Negative***
Positive	*TP*	*FP*
Negative	*FN*	*TN*

True Positive (TP) merupakan data positif yang terdeteksi benar, sedangkan *False Negative* (FN) adalah data positif namun terdeteksi sebagai data negatif. *True Negative* (TN) adalah jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Berdasarkan nilai TN, FP, FN, dan TP dapat diperoleh nilai akurasi, presisi dan recall (sensitifitas) [13].

Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar, dihitung dengan menggunakan persamaan (5). Presisi merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif, dihitung dengan menggunakan persamaan (6). *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif, dihitung dengan menggunakan persamaan (7) [13].

$$Accuracy = \frac{TP\ +\ TN}{TP + FP + FN + TN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

Keterangan :
- TP = Jumlah true *positive*
- TN = Jumlah true *negative*
- FP = jumlah false *positive*
- FN = jumlah false *negative*

Dalam mengevaluasi keunggulan model, perhitungan *Area Under Curve* (AUC) menjadi kunci. AUC mengukur luas area di bawah kurva pada grafik yang menggambarkan hubungan antara *sensitivity* dan *specificity*. Rentang nilai AUC antara 0 hingga 1, dengan nilai yang lebih tinggi menandakan kinerja yang lebih baik [14]. Terdapat kriteria interpretasi untuk nilai AUC, yaitu :
- Nilai AUC > 0,5 - 0,6: sangat lemah
- Nilai AUC > 0,6 - 0,7: lemah
- Nilai AUC > 0,7 - 0,8: sedang
- Nilai AUC > 0,8 - 0,9: baik
- Nilai AUC > 0,9 - 1: sangat baik

III. HASIL DAN PEMBAHASAN

A. Pemahaman Data (*Data Understanding*)

Dalam penelitian ini, digunakan data sekunder yang diperoleh melalui situs web *Kaggle.com* dengan judul "*Brain Stroke Prediction Dataset*". *Dataset* yang berhasil dikumpulkan terdiri dari 11 atribut dan total 4981 *record*. Dalam *dataset* ini, terdapat 10 atribut yang digunakan sebagai atribut reguler (*regular attributes*) dan satu atribut yang digunakan sebagai atribut khusus (*special attribute*). Nama dan keterangan setiap atribut dapat dilihat pada Tabel II.

TABEL II
ATRIBUT DAN TIPE ATRIBUT

No	**Nama *Attribut***	**Tipe *Atribut***
1	*Gender*	*Categorical*
2	*Age*	*Numeric*
3	*Hypertension*	*Categorical*
4	*Heart disease*	*Categorical*
5	*Ever Married*	*Categorical*
6	*Work Type*	*Categorical*
7	*Residence Type*	*Categorical*
8	*Average Glucose Level*	*Numeric*
9	*Body Mass Indeks (BMI)*	*Numeric*
10	*Smoking*	*Categorical*
11	*Stroke*	*Categorical*

B. Persiapan Data (*Data Preparation*)

1. Normalisasi

Normalisasi dilakukan mengubah representasi data dari satu bentuk ke bentuk lain yang lebih sesuai atau lebih mudah dipahami. Dalam penelitan ini diubah attribut yang memiliki

record dengan nilai 1 atau 0, dan kemudian nilai-nilai tersebut diubah menjadi *yes* atau *no* pada atribut *hypertension, heart_disease,* dan *stroke.*

2. *Discretize by Binning*

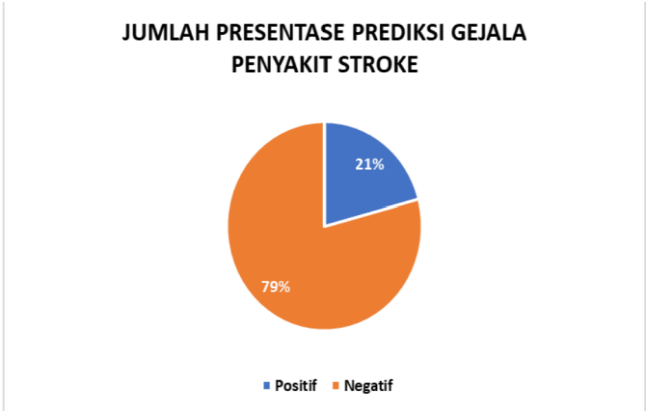
Discretization membagi data numerik ke dalam interval yang telah ditentukan dan mengubahnya menjadi atribut nominal. Dalam penelitian ini, digunakan dua interval (*number of bins* = 2) dalam proses *discretization*. Jumlah bin ialah jumlah rentang nilai pada atribut yang dituju yaitu *age* (umur). Pada atribut *age* jumlah *bin* adalah 2 sehingga memiliki 2 rentang pada atribut *age*. Kemudian dilakukan perbandingan antara dataset asli (tanpa proses diskritisasi) dan dataset hasil diskritisasi menggunakan algoritma DT, KNN, NB, RF dan SVM. Tabel III menunjukkan bahwa secara umum penerapan diskritisasi pada dataset meningkatkan akurasi model, kecuali pada model SVM.

TABEL III
PERBANDINGAN AKURASI PADA DATASET DENGAN DAN TANPA DISKRITISASI

Dataset	Akurasi Algoritma				
	DT	KNN	NB	RF	SVM
Tanpa Discretize	94,68%	94,22%	89,50%	94,96%	**92,81%**
Discretize	**94,84%**	**94,84%**	**90,32%**	**94,98%**	86,57%

3. *Data Imbalanced*

Dalam penelitian ini data yang digunakan cenderung tidak seimbang (*imbalanced*) dan memiliki kecenderungan ke arah kategori negatif. Pendekatan ini bertujuan untuk meningkatkan performa klasifikasi dengan menambahkan data sintetis (buatan) ke kelas minoritas agar seimbang dengan kelas mayoritas, dalam hal ini kelas positif. Penerapan SMOTE terhadap *dataset* hasil *preprocessing* menambahkan jumlah kelas minoritas sebanyak 1000 data, sehingga jumlah akhir *dataset* menjadi 5981 data dengan 1248 positif stroke dan 4733 negatif stroke. Jika dipresentasikan ke dalam bentuk diagram, terdapat 21% positif stroke dan 79% negatif stroke. Gambar 2 menyajikan komposisi data setelah dilakukan SMOTE.



Gambar 2. Komposisi Dataset setelah dilakukan SMOTE

4. Penentuan Data Latih dan Data Uji

Penentuan data latih dan data uji dilakukan dengan cara splitting data dan Cross Validation. Eksplorasi ini dilakukan untuk menentukan komposisi data dengan nilai AUC terbaik. Split Data dilakukan dengan komposisi data latih dan uji 80:20, 70:30 dan 60:40. Selain itu digunakan juga 10-fold Cross Validation, dimana data latih dan data uji dibagi kedalam 10 tempat atau partisi yang berbeda (*fold*) dan dilakukan 10 kali iterasi. Dataset asli dibandingkan dengan dataset hasil implementasi SMOTE. Algoritma yang digunakan dalam proses perbandingan adalah Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), dan Support Vector Machine (SVM). Hasil perbandingan menggunakan nilai AUC yang dapat dilihat pada Tabel IV. Penggunaan SMOTE terbukti dapat meningkatkan nilai AUC. Model terbaik dihasilkan dengan menggunakan algoritma *Random Forest* pada komposisi data latih dan data uji 60:40, dengan nilai AUC sebesar 0.836.

TABEL IV
PERBANDINGAN AUC PADA DATASET DENGAN ATAU TANPA SMOTE

	Ratio Perban-dingan	Algoritma				
		DT	KNN	NB	RF	SVM
Tanpa SMOTE	60:40	0.685	0.574	0.774	0.764	0.406
	70:30	0.659	0.549	0.773	0.748	0.738
	80:20	0.649	0.540	0.760	0.735	0.726
	Cross Validation	0.723	0.571	0.783	0.772	0.417
	Ratio Perbandin gan	Algoritma				
		DT	KNN	NB	RF	SVM
SMOTE	60:40	0.751	0.729	0.799	0.836	0.712
	70:30	0.771	0.741	0.793	0.828	0.484
	80:20	0.773	0.740	0.802	0.828	0.510
	Cross Validation	0.757	0.742	0.794	0.819	0.477

Dalam kasus ketidakseimbangan data (*imbalance dataset*), pemilihan model terbaik membutuhkan penggunaan nilai AUC (*Area Under the Curve*) sebagai metrik evaluasi yang lebih tepat daripada akurasi [7]. Nilai akurasi dapat dianggap kurang informatif dalam *imbalance dataset* karena cenderung mempelajari data mayoritas saja, mengabaikan data penting dari kelas minoritas dan berpotensi menyebabkan bias atau *overfitting*. Oleh karena itu, menggunakan nilai AUC memungkinkan evaluasi yang lebih komprehensif terhadap kinerja model dalam membedakan antara kedua kelas, terlepas dari ketidakseimbangan distribusi datanya.

C. Pemodelan

Berdasarkan hasil perbandingan akurasi dan AUC, maka Random Forest memiliki performa terbaik. Model Random Forest dibangun dengan number of trees 5 dan maximal depth 5 [15]. Berikut adalah aturan yang dihasilkan dari salah satu tree.

1) Jika Darah Tinggi (*Hypertension*) = yes dan massa indeks tubuh (BMI) <= 20.711 maka Stroke = yes
2) Jika Darah Tinggi (*Hypertension*) = yes dan massa indeks tubuh (BMI) > 20.711 dan massa indeks tubuh (BMI) <= 46.085 dan penyakit jantung (*Heart disease*) = yes, maka stroke = yes
3) Jika Darah Tinggi (*Hypertension*) = yes dan massa indeks tubuh (BMI) > 20.711 dan massa indeks tubuh (BMI) <= 46.085 dan penyakit jantung (*heart disease*) = no maka stroke = no
4) Jika Darah Tinggi (*Hypertension*) = yes dan massa indeks tubuh (BMI) > 20.711 dan massa indeks tubuh (BMI) > 46.085 maka stroke = yes
5) Jika Darah Tinggi (*Hypertension*) = no dan penyakit jantung (*Heart disease*) = yes dan kadar glukosa rata-rata (*average glucose level*) <= 108.041 maka stroke = no
6) Jika Darah Tinggi (*Hypertension*) = no dan penyakit jantung (*Heart disease*) = yes dan kadar glukosa rata-rata (*average glucose level*) > 108.041 dan pekerja wiraswasta (*self-employed*) = yes maka stroke = no
7) Jika Darah Tinggi (*Hypertension*) = no dan penyakit jantung (*Heart disease*) = yes dan kadar glukosa rata-rata (*average glucose level*) > 108.041 dan pekerja swasta (*private job*) = yes maka stroke = yes
8) Jika Darah Tinggi (*Hypertension*) = no dan penyakit jantung (*Heart disease*) = yes dan kadar glukosa rata-rata (*average glucose level*) > 108.041 dan pekerja pemerintah (*govt job*) = yes maka stroke = yes
9) Jika Darah Tinggi (*Hypertension*) = no dan penyakit jantung (*Heart disease*) = no maka stroke = no

D. Evaluasi

Pada tahap evaluasi dilakukan pengujian algoritma dengan model terbaik menggunakan *confusion matrix*. Parameter uji yang digunakan adalah akurasi, presisi dan recall yang dihitung menggunakan persamaan (5), (6), dan (7). Tabel V menyajikan confusion matrix dari model terbaik, yaitu model SMOTE dan Random Forest dengan komposisi dataset 60:40.

TABEL V
CONFUSION MATRIX

	True Yes	*True No*	*Class Precission*
Prediction Yes	57	33	63.33%
Prediction No	442	1860	80.80%
Class Recall	11.42%	98.26%	

Dalam penelitian ini penggunaan *Synthetic Minority Over Sampling Technique* (SMOTE) dalam pemodelan *Random Forest* terbukti efektif dalam meningkatkan nilai AUC namun tidak terjadi peningkatan pada akurasi. Peningkatan nilai AUC pada pemodelan *Random Forest* menggunakan *split data* pada komposisi data latih 60:40 data uji cukup signifikan yaitu, dari 0.764 tanpa menggunakan SMOTE menjadi 0.836 dengan penggunaan SMOTE yang mengalami kenaikan sebesar 0.072.

IV. KESIMPULAN

Berdasarkan hasil pengujian dan analisis yang dilakukan serta perbandingan untuk mencari model terbaik, penggunaan Teknik SMOTE dan algoritma Random Forest dalam penelitian ini terbukti dapat perfoma yang baik pada data tidak seimbang serta meningkatkan nilai AUC namun tidak meningkatkan nilai akurasi pada dataset penelitian ini. Total data pada penelitian ini sebanyak 4981 record dengan jumlah negatif sebanyak 4733 atau 95% dan positif sebanyak 248 atau 5%. Setelah dilakukannya teknik SMOTE terbentuk sebanyak 5981 record dengan jumlah positif 1.248 atau sekitar 21% dan 4733 atau sekitar 79% jumlah negatif. Metode validasi yang digunakan pada penelitian ini adalah split data pada komposisi data latih 60:40 data uji menggunakan algoritma *Random Forest* dengan SMOTE menghasilkan akurasi sebesar 80,14% dan AUC 0,0836.

REFERENSI

[1] Md. M. Islam, et al., 'Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique', *International Journal of Electronics and Communications Systems*, vol. 1, no. 2, pp. 57–62, 2021, doi: 10.24042/ijecs.v1i2.10393.
[2] S. Siswanto, 'Laporan Nasional RISKESDAS 2018', *Kementerian Kesehatan RI*, vol. 1, no. 1, p. 1, 2019.
[3] A. F. Riany, and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *Jurnal Saintekom: Sains, Teknologi, Komputer dan Manajemen*, vol. 13, no. 1, pp. 42–54, 2023, https://doi.org/10.33020/saintekom.v13i1.352.
[4] M. N. Maskuri, H. Harliana, K. Sukerti, and R. M. H. Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm," *Jurnal Ilmiah Intech: Information Technology Journal of UMUS*, vol. 4, no. 1, pp. 130–140, 2022, https://doi.org/10.46772/intech.v4i01.751.
[5] A. F. Hermawan, F. R. Umbara, and F. Kasyidi, "Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis menggunakan Metode Algoritma CART (Classification and Regression Tree)," *MIND (Multimedia Artificial Intelligent Networking Database) Journal*, vol. 7, no. 2, pp. 151–164, 2022, https://doi.org/10.26760/mindjournal.v7i2.151-164.
[6] R. E. Pambudi, S. Sriyanto, and F. Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C4.5," *Jurnal Teknika*, vol. 16, no. 2, pp. 221–226, 2022, https://doi.org/10.5281/zenodo.7535865.
[7] M. I. Putri and I. Kharisudin, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Analisis Sentimen Data Review Pengguna Aplikasi Marketplace Tokopedia," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 5, pp. 759–766, 2022.
[8] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Jurnal of Artificial Intelligence*, vol. 16, pp. 321-357, 2002, https://doi.org/10.1613/jair.953.
[9] A. Rohman and M. Rochcham, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Mahasiswa," *Neo Teknika*, vol. 5, no. 1, pp. 23–29, 2019, doi: 10.37760/neoteknika.v5i1.1379.
[10] B. Budiman, "Perbandingan Algoritma Klasifikasi Data Mining untuk Penelusuran Minat Calon Mahasiswa Baru," *Nuansa Informatika*, vol. 15, no. 2, pp. 37–52, 2021, doi: 10.25134/nuansa.v15i2.4162.
[11] Luthfiana Ratnawati and Dwi Ratna Sulistyaningrum, 'Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel', *Jurnal Sains Dan Seni ITS*, vol. 8, no. 2, pp. 71–77, 2019.
[12] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, pp. 208-214, 2022, doi: 10.30865/mib.v6i1.3373.
[13] J. Han, M. Kamber, and J. Pei, *Data Mining Concept and Techniques*, 3rd ed. USA: Morgan Kaufmann Publishers, 2012.
[14] Qadrini L, Sepperwali A, and Aina A, "Decision Tree dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial," *Jurnal Inovasi Penelitian*, vol. 2, no. 7, pp. 1959–1966, 2021, https://doi.org/10.47492/jip.v2i7.1046.

[15] A. C. Mawarni, R. Rusdah, L. L. Hin, and D. Anubhakti, 'Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest', *IDEALIS : InDonEsiA journaL Information System*, vol. 6, no. 2, pp. 165–171, Jul. 2023, doi: 10.36080/idealis.v6i2.3018.