Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 30-38

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

PERBANDINGAN KINERJA TEKNIK PENYEIMBANG KUMPULAN DATA DALAM MODEL PREDIKSI PENYAKIT STROKE

Eugene Vincent Arends
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *eugene.535210082@stu.untar.ac.id*

ABSTRAK

Stroke adalah salah satu penyakit neurologis yang berbahaya dimana pembuluh darah otak tersumbat ataupun mengalami pemecahan. Prediksi *stroke* dengan akurat dapat membantu perawatan dan penanganan penyakit *stroke*. Model prediksi dapat dibuat dengan teknik pembelajaran mesin dengan menggunakan data pasien dari rumah sakit. Dalam data pra-proses, semua data dengan nilai kosong dihapus, dan teknik penyeimbangan dataset yang digunakan adalah *Random Oversampling* (ROS), *Random Undersampling* (RUS), dan *Synthetic Minority Oversampling Technique* (SMOTE). Model pembelajaran mesin yang digunakan dalam tahap eksperimen adalah *Gaussian Naïve Bayes* (GNB), *Logistic Regression* (LR), dan *K-Nearest Neighbors* (K-NN). *Sensitivity* sebelum menggunakan teknik *balancing* adalah 33% untuk GNB dan LR, dan 1% untuk K-NN. Setelah menggunakan teknik *balancing*, nilai *sensitivity* terbesar untuk GNB dan LR adalah 76%, dan nilai sensitivity terbesar untuk K- NN adalah 73%. Dapat disimpulkan bahwa teknik penyeimbangan kumpulan data memainkan peran besar dalam meningkatkan performa model pada kumpulan data yang sangat tidak seimbang, seperti kumpulan data yang digunakan dalam penelitian ini.

Kata kunci: *Random Oversampling, Random Undersampling, Stroke, SMOTE.*

ABSTRACT

Stroke is a dangerous neurological disease in which the blood vessels in the brain become blocked or rupture. Accurate stroke predictions can help treat stroke and improve care. Prediction models can be created using machine learning techniques using patient data from hospitals. During the pre-processing stage, data with empty values are deleted, and the balancing techniques used are Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Oversampling Technique (SMOTE). The machine learning models used in the experiment are Gaussian Naïve Bayes (GNB), Logistic Regression (LR), and K-Nearest Neighbors (K-NN). The Sensitivity of the model is the evaluation method that is valued in this research. Sensitivity score before using the balancing technique is 33% for GNB and LR, and 1% for K-NN. After using the balancing technique, the greatest sensitivity score for GNB and LR was 76%, and for K-NN it was 73%. It can be concluded that dataset balancing techniques play a big role in improving model performance on highly imbalanced datasets, such as the dataset used in this research.

Keywords: *Random Oversampling, Random Undersampling, Stroke, SMOTE.*

1. PENDAHULUAN

Stroke adalah sebuah penyakit yang disebabkan oleh penyumbatan pembuluh darah. Pecahnya arteri yang menuju ke otak sebagai akibat dari penyumbatan pembuluh darah mengakibatkan kematian sel-sel otak akibat kekurangan oksigen. Peradangan saraf yang terjadi dapat mempengaruhi pemulihan dan komplikasi pasca *stroke* [1]. Kelainan saraf dan gangguan pencernaan merupakan beberapa contoh dampak sekunder akibat perubahan hormon yang terjadi akibat *stroke* [2]. Penyakit ini merupakan penyebab kematian nomor dua di dunia dan penyebab kematian dan disabilitas nomor tiga di dunia bila digabungkan [3]. *Stroke* dapat timbul pada siapapun, namun lebih sering pada usia tua karena degradasi pada dinding pembuluh darah [4]. Selain itu, *stroke* juga lebih sering timbul dan lebih berdampak secara negatif terhadap wanita [5].

Disabilitas jangka panjang akibat *stroke* dapat mengganggu kemampuan melakukan aktivitas sehari-hari atau kembali ke dunia kerja yang mengurangi rasa harga diri [6]. Pemulihan disabilitas

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

fisik pasca *stroke* melalui terapi fisik baik dalam rumah sakit ataupun lingkungan rumah [7]. Untuk mengurangi beban *stroke*, strategi pencegahan *stroke* sangat penting, terutama dengan pertumbuhan populasi dan penuaan [8]. Salah satunya adalah prediksi *stroke* dengan menggunakan pembelajaran mesin. Komputer diberikan data pasien dan membangun sebuah model prediksi dengan algoritma yang ditentukannya.

2. TINJAUAN LITERATUR

Berbagai penelitian telah dilakukan untuk mengevaluasi algoritma yang terbaik untuk memprediksi stroke. Emon et al. [9], Heo et al. [10] dan Shoily et al. [11] membandingkan berbagai algoritma dan menemukan akurasi setinggi 97% dengan Weighted Voting Classifier [9]. Dev et al. menemukan bahwa algoritma Neural Network dengan fitur umur, riwayat kelainan jantung, rata-rata kadar glukosa, dan riwayat hipertensi dengan akurasi 78% dan miss rate 19% [12]. Beberapa penelitian menerapkan balancing technique untuk menangani kumpulan data yang tidak seimbang, yang ditunjukkan dalam Tabel 1.

Tabel 1 *Balancing technique* yang digunakan oleh beberapa penelitian terkait

Penelitian	*Balancing technique(s)* yang digunakan
Sailasya et al. [13]	*Random Undersampling*
Dristas et al. [14]	*Synthetic Minority Oversampling Technique*
Nwosu et al. [15]	*Random Undersampling*
Tazin et al. [16]	*Synthetic Minority Oversampling Technique*
Biswas et al. [17]	*Random Oversampling*

Dalam penelitian ini, *Random Undersampling* (RUS), *Random Oversampling* (ROS), dan *Synthetic Minority Oversampling Technique* (SMOTE) adalah tiga *balancing technique* yang kinerjanya akan dibanding dengan model klasifikasi stroke tanpa *balancing technique*. Tiga algoritma klasifikasi yang akan digunakan untuk penelitian ini adalah *Gaussian Naïve Bayes* (GNB), *Logistic Regression* (LR), dan *K-Nearest Neighbors* (K-NN). Penelitian ini diharapkan akan menunjukkan kinerja klasifikasi model klasifikasi stroke dengan dan tanpa menerapkan *balancing technique*..

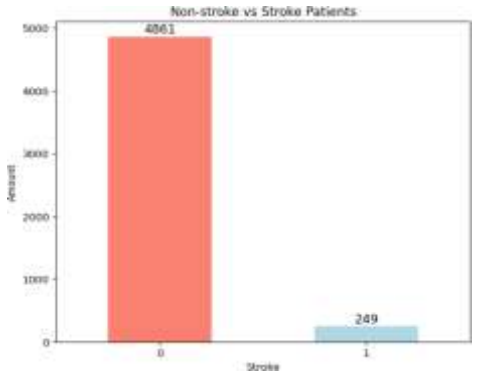
3. METODE PENELITIAN

3.1 Dataset

Kumpulan data yang digunakan dalam eksperimen ini berasal dari https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Dataset ini terdiri dari 5110 data pasien dengan 12 atribut dan 2 kelas: *stroke* dan *non-stroke*. Fitur yang dipilih adalah jenis kelamin, usia, menderita hipertensi, menderita penyakit jantung, status pernikahan, tipe pekerjaan, tipe tempat tinggal, kadar glukosa dalam darah, indeks massa tubuh, dan status merokok.

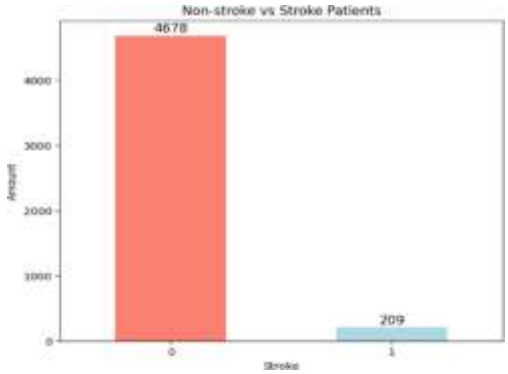
3.2 Pra-pemrosesan Data

Sebelum data dapat dilatih, data harus diolah agar dapat digunakan oleh model dalam tahap pra-pemrosesan. Atribut "*id*' dihapus karena tidak berhubungan dengan stroke dan dapat mengurangi kualitas model, dan semua atribut dalam format string disubstitusikan dengan sebuah nilai numerik. 4861 dari pasien dari dataset tidak menderita *stroke*, dan 249 pasien menderita *stroke*. Kumpulan data terdapat data dengan nilai yang kosong. Data tersebut dibersihkan dengan menghapus data pasien dengan kekosongan nilai. Data yang bersih ini terdapat 4678 pasien yang tidak menderita *stroke* dan 209 pasien yang menderita *stroke*. Pada Gambar 1 ditampilkan distribusi perbandingan pasien tidak *stroke* (0) dan pasien *stroke* (1).



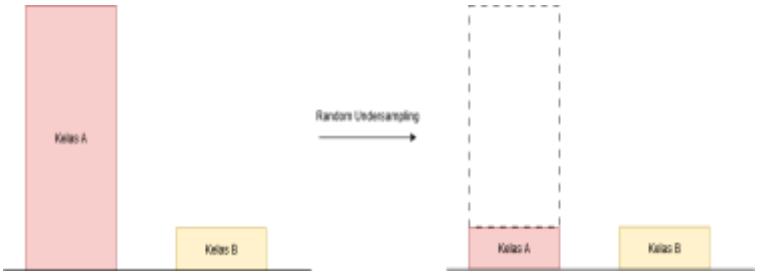
Gambar 1. Perbandingan data pasien tidak *stroke* (0) dan *stroke* (1) dalam dataset

Setelah dibandingkan, ternyata pada kelas target terdapat perbandingan antara pasien tidak terkena stroke dan pasien yang terkena stroke adalah 23:1. Ketidakseimbangan tersebut disebut sebagai *class imbalance*. Fenomena *class imbalance* ini dapat memunculkan hasil yang tidak baik untuk prediksi kelas minoritas pada model prediksi, yaitu untuk yang terkena stroke. Tiga balancing technique diterapkan untuk mengatasi masalah tersebut: *Random Undersampling* (RUS), *Random Oversampling* (ROS), dan *Synthetic Minority Oversampling Technique* (SMOTE). Pada Gambar 2 ditampilkan perbandingan data pasien tidak *stroke* (0) dan pasien *stroke* (1) dalam dataset yang sudah dibersihkan.



Gambar 2. Perbandingan data pasien tidak *stroke* (0) dan *stroke* (1) dalam dataset yang sudah dibersihkan

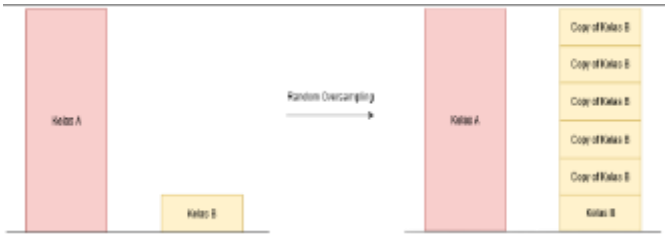
Kumpulan data yang tidak seimbang adalah kumpulan data yang memiliki jenis distribusi data yang mendominasi ruang instance secara signifikan dibandingkan distribusi data lainnya [18]. *Random Undersampling* **(**RUS) adalah *balancing technique* yang digunakan dalam tahap pra-pemrosesan yang mengurangi data yang berasal dari kelas mayoritas secara acak sehingga distribusi data antara semua kelas sama rata. Kekurangan dari teknik ini adalah kemungkinan untuk menghapus data yang penting untuk model pengklasifikasi [19]. Pada Gambar 3 ditunjukkan contoh diagram dari RUS.



Gambar 3. Diagram *Random Undersampling*

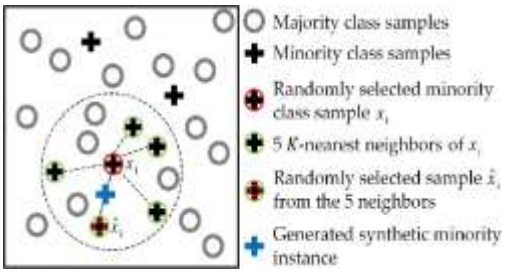
ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Oversampling adalah *balancing technique* yang menambahkan data kelas minoritas dalam sebuah *dataset* [18]. *Random Oversampling* (ROS) adalah metode *Oversampling* yang menyeimbangkan penyebaran kelas dengan menduplikasi secara acak instansi kelas minoritas. Kelemahan dari teknik ini adalah meningkatnya kemungkinan terjadinya *overfitting*, sebuah fenomena dimana model bekerja dengan baik dalam data testing namun bekerja dengan buruk untuk data eksternal, dan juga meningkatkan lamanya *training* model, terutama untuk kumpulan data yang sangat besar. Pada Gambar 4 ditunjukkan contoh diagram dari ROS.



Gambar 4. Diagram *Random Oversampling*

Synthetic Minority Oversampling Technique SMOTE salah satu *balancing technique* yang populer digunakan untuk kumpulan data yang tidak seimbang. SMOTE adalah pengembangan dari *Oversampling* dan ROS yang dikembangkan oleh Chawla et al. pada tahun 2002 sebagai sebuah pendekatan baru untuk menangani *imbalanced dataset* [20]. Setiap sampel kelas minoritas diambil dan dikenalkan contoh sintetik di sepanjang segmen garis yang menghubungkan salah satu atau semua *k-neighbors* terdekat kelas minoritas [21]. Kelemahan teknik ini adalah *sample overlap* dan *noise interference*.



Gambar 5. Diagram SMOTE (Sumber: *Rich Data*)

3.3 Algoritma Klasifikasi

Algoritma klasifikasi digunakan untuk memprediksi pasien stroke atau tidak stroke. Algoritma harus dipilih dengan baik dan yang cocok dengan kumpulan data yang ada. Algoritma yang digunakan dalam penelitian ini adalah: *Gaussian Naïve Bayes*, *Logistic Regression*, dan *K-Nearest Neighbors*.

Gaussian Naïve Bayes (GNB) adalah sebuah teknik klasifikasi berdasarkan teorema *Bayes* dan distribusi *Gaussian* (distribusi normal). Dalam teknik ini, GNB mengasumsi bahwa setiap fitur mengikuti distribusi normal (*Gaussian*) dan bahwa setiap fitur tidak berhubungan satu sama lain (Naïve Bayes) [22]. Persamaan matematis dari algoritma ini dapat dinyatakan dalam rumus sebagai berikut.

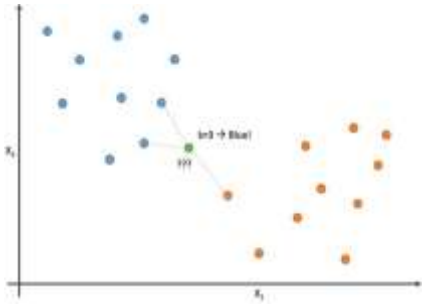
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}$$

Logistic Regression (LR) adalah sebuah teknik klasifikasi yang sangat populer dan banyak digunakan untuk berbagai aplikasi. Dengan menggunakan regresi logistik, sebuah hasil variable antara 0 dan 1 didapatkan, dan dengan fungsi sigmoid, menghasilkan sebuah hasil prediksi biner [16].

K-Nearest Neighbors (K-NN) adalah sebuah teknik klasifikasi yang menggunakan *Euclidean distance* (2) untuk membuat sebuah prediksi atau klasifikasi. Nilai *k* dalam penelitian ini

adalah *5*. *K-value* yang dipilih akan menentukan jumlah tetangga yang akan diperhatikan dalam prediksi. Persamaan algoritma ini yang dapat dinyatakan sebagai berikut dan Gambar 6 menunjukkan contoh diagram *k-nearest neighbors*.

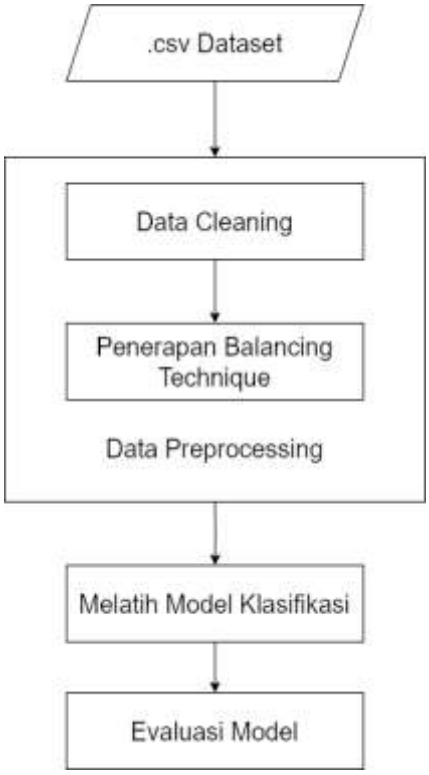
$$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$



Gambar 6. Diagram *K-Nearest Neighbors* (Sumber: DiSHA Computer Institute)

3.4 Rancangan Eksperimen

Alur eksperimen dalam penelitian ini digambarkan pada Gambar 7. Eksperimen ini dimulai dengan dataset *stroke* dalam format *csv* yang dibersihkan dan diterapkan tiga teknik keseimbangan dataset, yang digunakan untuk melatih tiga model klasifikasi yang akan dievaluasi.



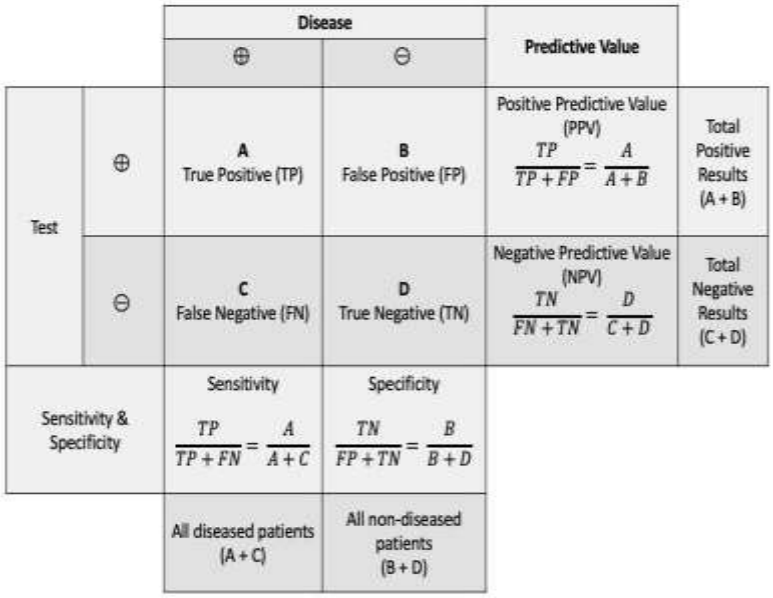
Gambar 7. Alur Kerja Penelitian dan Evaluasi Model

3.4 Rancangan Eksperimen

Metode evaluasi yang dilakukan adalah mencari nilai *sensitivity*, *specificity* dan akurasi dari setiap model dan dibandingkan. Nilai sensitivitas, spesifisitas, dan akurasi lebih tinggi, semakin baik model tersebut. Evaluasi *sensitivity* dipilih karena dataset yang sangat tidak seimbang. Ketidakseimbangan ini dapat menghasilkan *overfitting* pada kelas mayoritas dan *underfitting* pada

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

kelas minoritas. Dengan ketiganya, model akan terevaluasi dari kelas mayoritas dan minoritas. Pada Gambar 8 berikut ini ditunjukkan sensitivitas dan spesifisitas dalam metode evaluasi yang diterapkan dalam penelitian ini.



Gambar 8. *Sensitivity* dan *Specificity* (Sumber: Medbullets)

4. HASIL DAN PEMBAHASAN

Dataset dipisah menjadi data latih dan data uji dengan perbandingan 7:3. Model GNB, LR, dan K-NN dilatih dengan data tersebut. Akurasi ketiga model sangat baik, namun, nilai *sensitivity* kurang baik, dan terutama sangat buruk pada model K-NN, seperti ditunjukkan oleh Tabel 2.

Tabel 2. Hasil Evaluasi Ketiga Model untuk Data Tanpa *Balancing*

Model	Akurasi	*Specificity*	*Sensitivity*
GNB	86%	0,8846	0,3333
LR	86%	0,8846	0,3333
K-NN	95%	0,9936	0,0159

Model dilatih dengan data yang sama, tetapi diseimbangkan dengan *Random Undersampling*. Model yang telah dilatih dievaluasi dengan data uji yang sama. Seperti ditunjukkan oleh Tabel 3, akurasi dan *specificity* untuk ketiga *classifier* menurun sedikit, namun *sensitivity* meningkat pesat.

Tabel 3. Hasil Evaluasi Ketiga Model setelah Menerapkan RUS

Model	Akurasi	*Specificity*	*Sensitivity*
GNB	75%	0,7557	0,6825
LR	75%	0,7557	0,6825
K-NN	73%	0,7258	0,7302

Data diseimbangkan dengan *Random Oversampling* dan dievaluasi dengan metode yang sama. Seperti ditunjukkan oleh Tabel 4, akurasi dan *specificity* untuk ketiga *classifier* menurun sedikit, *sensitivity* untuk GNB dan LR meningkat setara dengan model GNB dan LR dengan penerapan RUS, sedangkan *sensitivity* K-NN dengan RUS hanya meningkat sangat sedikit.

Tabel 4. Hasil Evaluasi Ketiga Model setelah Menerapkan ROS

Model	Akurasi	*Specificity*	*Sensitivity*
GNB	75%	0,7528	0,6825
LR	75%	0,7557	0,6825
K-NN	87%	0,9003	0,2857

Data diseimbangkan dengan *Synthetic Minority Oversampling Technique* dan dievaluasi dengan metode yang sama. Seperti ditunjukkan oleh Tabel 5, akurasi dan *specificity* untuk ketiga *classifier* menurun sedikit, *sensitivity* untuk GNB dan LR meningkat setara dengan model GNB dan LR dengan penerapan RUS, sedangkan *sensitivity* K-NN dengan RUS hanya meningkat sangat sedikit.

Tabel 5. Hasil Evaluasi Ketiga Model setelah Menerapkan SMOTE

Model	Akurasi	*Specificity*	*Sensitivity*
GNB	64%	0,6303	0,7619
LR	64%	0,7557	0,6825
K-NN	80%	0,8113	0,5079

5. KESIMPULAN

Dalam penelitian ini, kinerja tiga algoritma pengklasifikasi dan tiga *balancing technique* untuk kumpulan data pasien stroke dengan kelas target yang tidak seimbang. Akurasi, *sensitivity,* dan *specificity* adalah ukuran kinerja setiap algoritma.

Dari Tabel 6, semua *classifier* bekerja berbeda dengan setiap *balancing technique.* K-NN dengan *k-value* 5 tanpa *balancing technique* memiliki hasil dengan akurasi dan *specificity* terbaik, tetapi dengan *sensitivity* terendah. Model ini menghasilkan prediksi 1395 dari 1404 *non-stroke* dan 1 dari 63 stroke, yang menunjukkan fenomena *underfitting*. Sebaliknya, GNB dan LR dengan SMOTE menunjukkan *sensitivity* terbaik dengan akurasi dan *specificity* terendah.

Model ini menghasilkan prediksi 885 dari 1404 *non-stroke* dan 48 dari 63 stroke, yang menunjukkan fenomena *overfitting,* yang ditunjukkan oleh akurasi prediksi kelas mayoritas (*non-stroke*) yang lebih tinggi dibandingkan dengan akurasi prediksi kelas minoritas (stroke). Dari hasil yang sangat beragam, dapat disimpulkan bahwa untuk *dataset* yang *imbalanced, balancing technique* dapat meningkatkan kualitas model prediksi. Secara praktis, setiap *classifier* dan *balancing technique* harus disesuaikan dengan *dataset* yang ada.

Tabel 6. Hasil Akurasi, *Specificity,* dan *Sensitivity* untuk Setiap Skenario Penelitian

Model	Akurasi	*Specificity*	*Sensitivity*
GNB	86%	0,8846	0,3333
LR	86%	0,8846	0,3333
K-NN	95%	0,9936	0,0159
GNB + RUS	75%	0,7557	0,6825
LR + RUS	75%	0,7557	0,6825
K-NN + RUS	73%	0,7258	0,7302
GNB + ROS	75%	0,7528	0,6825
LR + ROS	75%	0,7528	0,6825
K-NN + ROS	87%	0,9003	0,2857
GNB + SMOTE	64%	0,6303	0,7619
LR + SMOTE	64%	0,6303	0,7619
K-NN + SMOTE	80%	0,8113	0,5079

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

DAFTAR PUSAKA

[1] . Kuriakose and Z. Xiao, "Pathophysiology and Treatment of Stroke: Present Status and Future erspectives," *International Journal of Molecular Sciences,* vol. 21, no. 20, 2020.

[2] . Ge, M. Zadeh, C. Yang, E. Candelario-Jalil and M. Mohamadzadeh, "Ischemic Stroke Impacts e Gut Microbiome, Ileal Epithelial and Immune Homeostasis," *iScience,* vol. 25, no. 11, 2022.

[3] . V. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian ıd P. Lindsay, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *International ɔurnal of Stroke,* vol. 17, no. 1, pp. 18-29, 2022.

[4] . Young and M. Yousufuddin, "Aging and ischemic stroke," *Aging (Albany NY),* vol. 11, no. 9, pp. 542-2544, 2019.

[5] . M. Rexrode, T. E. Madsen, A. Y. X. Yu, C. Carcel, J. H. Lichtman and E. C. Miller, "The Impact Sex and Gender on Stroke," *Circulation Research,* vol. 130, no. 4, pp. 512- 528, 2022.

[6] . M. Martin-Saez and N. James, "The experience of occupational identity disruption post stroke: systematic review and meta-ethnography," *Disability and Rehabilitation,* vol. 43, no. 8, pp. 1044-)55, 2021.

[7] .-C. Chiu, F.-C. Chi and P.-T. Chen, "Investigation of the home-reablement program on habilitation outcomes for people with stroke," *Medicine,* vol. 100, no. 26, 2021.

[8] B. Gorelick, "The global burden of stroke: persistent and disabling," *The Lancet Neurology,* ɔl. 18, no. 5, pp. 417-418, 2019.

[9] . U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, S. M. Kaiser and S. A. Mamun, Performance Analysis of Machine Learning Approaches in Stroke Prediction," in *4th IEEE ıternational Conference on Electronics, Communication and Aerospace Technology (ICECA)20),* Coimbatore, 2020.

[10] Heo, G. J. Yoon, Y. D. Kim, H. S. Nam and J. H. Heo, "Machine Learning–Based Model for rediction of Outcomes in Acute Stroke," *Stroke,* p. 1263–1265, 2019.

[11] . I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif and R. R. Ema, "Detection of Stroke Disease sing Machine Learning Algorithms," in *10th International Conference on Computing, ommunication and Networking Technologies (ICCCNT),* Kanpur, 2019.

[12] Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli and D. John, "A predictive analytics approach ır stroke prediction using machine learning and neural networks," *Healthcare Analytics,* vol. 2,)22.

[13] . Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML lassification Algorithms," *International Journal of Advanced Computer Science and Applications JACSA),* vol. 12, no. 6, pp. 539-545, 2021.

[14] . Dristas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *ensors,* vol. 22, no. 13, 2022.

[15] . S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli and D. John, "Predicting Stroke from Electronic ealth Records," in *Annual International Conference of the IEEE Engineering in Medicine and iology Society (EMBC),* Glasgow, 2019.

[16] . Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis and M. M. Khan, "Stroke Disease Detection ıd Prediction Using Robust Learning Approaches," *Journal of Healthcare Engineering,* 2021.

[17] . Biswas, K. M. M. Uddin, S. T. Rikta and S. K. Dey, "A comparative analysis of machine learning assifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics,* vol. 2,)22.

[18] I. S. Shelke, P. R. Deshmukh and V. K. Shandilya, "A Review on Imbalanced Data Handling sing Undersampling and Oversampling Technique," *International Journal of Recent Trends in ngineering and Research,* vol. 3, no. 4, 2017.

[19] . Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and ndersampling Techniques: Overview Study and Experimental Results," in *20th International onference on Information and Communications System (ICICS),* Copenhagen, 2020.

[20] . Fernández, S. Gárcia, F. Herrera and N. V. Chawla, "SMOTE for Learning from Imbalanced ata: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence esearch,* vol. 61, pp. 863-905, 2018.

[21] . V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority ver-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[22] . Pundlik, "Comparison of Sensitivity for Consumer Loan Data Using Gaussian Naïve Bayes GNB) dan Logistic Regression (LR)," in *7th International Conference on Intelligent Systems, Modelling and Simulatio*, Bangkok, 2016.

[23] . Wu and Y. Fang, "Stroke Prediction with Machine Learning Methods among Older Chinese," *nternational Journal of Environmental Research and Public Health,* vol. 17, no. 6, p. 1828, 2020.

[24] . Prentzas, C. S. Pattichis and A. C. Kakas, "Integrating Machine Learning with Symbolic easoning to Build an Explainable AI Model for Stroke Prediction," in *2019 IEEE 19th nternational Conference on Bioinformatics and Bioengineering (BIBE),* Athens, 2019.

[25] Paliwal, S. Parveen, M. A. Alam and J. Ahmed, "Improving Brain Stroke Prediction through versampling Techniques: A Comparative Evaluation of Machine Learning Algorithms," *reprints,* 2023.

[26] . Y. Boateng, J. Otoo and D. A. Abaye, "Basic Tenets of Classification Algorithms K- Nearest-eighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal f Data Analysis and Information Processing,* vol. 8, no. 4, pp. 341-357, 2020.