HEXATECH
Jurnal Ilmiah Teknik
Vol 4 No 2 Agustus 2025
ISSN: 2828-8696 (Print) ISSN: 2828-8548 (Electronic)
Open Access: https://jurnal.arkainstitute.co.id/index.php/hexatech/index

Optimalisasi prediksi penyakit stroke pada data tidak seimbang menggunakan teknik SMOTE-Tomek

Eka Herdit Juningsih[1], Kartika Handayani[2], Erni[3], Dinar Ismunandar[4]
[1,2,3,4]Universitas Bina Sarana Informatika
email: [1]eka.ekj@bsi.ac.id, [2]kartika.kth@bsi.ac.id, [3]erni.erx@bsi.ac.id, [4]dinar.dim@bsi.ac.id

Info Artikel :
Diterima :
13 Juni 2025
Disetujui :
10 Juli 2025
Dipublikasikan :
25 Juli 2025

ABSTRAK

Stroke merupakan salah satu penyakit yang berdampak besar terhadap kesehatan masyarakat, karena dapat menyebabkan kematian maupun kecacatan permanen. Oleh karena itu, identifikasi dini terhadap risiko stroke menjadi sangat penting dalam upaya pencegahan. Penelitian ini dilakukan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu *Naive Bayes*, *K-Nearest Neighbors* (KNN), dan *Decision Tree*, dalam memprediksi kejadian stroke berdasarkan data rekam medis pasien. Dataset yang digunakan terdiri dari 5.110 entri dengan sejumlah fitur seperti usia, jenis kelamin, tekanan darah, kadar glukosa, dan status merokok. Teknik *SMOTE-Tomek* diterapkan untuk menyeimbangkan jumlah antara data pasien stroke dan non-stroke. Selanjutnya, model dievaluasi menggunakan pengujian validasi silang lima lipatan dan diukur menggunakan akurasi, *precision*, *recall*, serta *f1-score*. Hasil pengujian menunjukkan bahwa model *Decision Tree* memberikan performa terbaik dengan nilai f1-score yang paling tinggi yakni 91%, setelah penyeimbangan data dilakukan. Berdasarkan hasil pengujian, algoritma *Decision Tree* menunjukkan performa terbaik dengan nilai *f1-score* sebesar 91% setelah penyeimbangan data menggunakan SMOTE-Tomek. Hal ini membuktikan bahwa kombinasi teknik *balancing* SMOTE-Tomek dan pemilihan algoritma klasifikasi yang sesuai sangat efektif dalam meningkatkan akurasi prediksi resiko stroke pada data tidak seimbang.

Kata kunci: Prediksi Stroke, Data Tidak Seimbang, SMOTE-Tomek, Algoritma Klasifikasi

ABSTRACT

Stroke is one of the diseases that has a significant impact on public health, as it can cause death or permanent disability. Therefore, early identification of stroke risk is very important in prevention efforts. This study was conducted to compare the performance of three classification algorithms—Naive Bayes, K-Nearest Neighbors (KNN), and Decision Tree—in predicting stroke occurrence based on patient medical records. The dataset used consisted of 5,110 entries with various features such as age, gender, blood pressure, glucose levels, and smoking status. The SMOTE-Tomek technique was applied to balance the number of stroke and non-stroke patient data. The models were then evaluated using five-fold cross-validation testing and measured using accuracy, precision, recall, and F1-score. The test results showed that the Decision Tree model performed best with the highest F1-score of 91% after data balancing was performed. Based on the testing results, the Decision Tree algorithm demonstrated the best performance with an F1-score of 91% after data balancing using SMOTE-Tomek. This proves that the combination of the SMOTE-Tomek balancing technique and the selection of an appropriate classification algorithm is highly effective in improving the accuracy of stroke risk prediction in imbalanced data.

Keywords : Stroke Prediction, Imbalanced Data, SMOTE-Tomek, Classification Algorithms

PENDAHULUAN

Stroke merupakan salah satu penyakit kardiovaskular dan neurologis yang paling sering dijumpai di kawasan Asia, termasuk di Indonesia (Turana et al., 2021). Di tingkat nasional, stroke tercatat sebagai penyebab utama kematian serta salah satu faktor terbesar penyebab kecacatan jangka panjang. Berdasarkan data terkini, prevalensi stroke di Indonesia mencapai 10,9%, dengan tingkat kejadian yang cenderung stabil, yaitu sekitar 120 kasus per 100.000 penduduk setiap tahunnya sejak tahun 2013 (Anisah Makkiyah, 2024). Kondisi ini tidak hanya berdampak pada aspek kesehatan individu, tetapi juga memberikan tekanan ekonomi yang signifikan terhadap rumah tangga. Beban finansial yang ditimbulkan oleh penyakit stroke diketahui lebih tinggi dibandingkan penyakit tidak menular lainnya, dengan estimasi pengeluaran mencapai sekitar 10,7% dari total pendapatan rumah tangga untuk

pembiayaan pengobatan dan perawatan penderita stroke (Riyadina, W., Pradono, J., Kristanti, D., & Turana, 2020). Stroke, atau yang dalam istilah medis dikenal sebagai *Cerebrovascular disease*, merupakan kondisi gangguan serius yang terjadi pada fungsi otak akibat gangguan pada aliran darah. Gangguan ini menyebabkan bagian tubuh tertentu mengalami kelumpuhan, baik sebagian maupun seluruh bagian tubuh, sehingga menghambat fungsi normal dari sistem gerak maupun sistem sensorik tubuh(Chavda et al., 2021).

Secara lebih rinci, stroke dapat diartikan sebagai suatu kondisi medis yang terjadi ketika suplai darah yang membawa oksigen dan nutrisi penting ke jaringan otak terhambat atau terganggu. Gangguan ini bisa disebabkan oleh penyumbatan (iskemik) atau pecahnya pembuluh darah (hemoragik) di otak. Ketika aliran darah terganggu, sel-sel otak yang tidak mendapatkan oksigen dan nutrisi akan mulai mati dalam waktu singkat. Proses ini disebut dengan infark serebral, yaitu kematian jaringan otak yang bersifat permanen akibat kurangnya pasokan darah yang cukup. Dampaknya bisa sangat luas, mulai dari kehilangan kemampuan bicara, gangguan memori, hingga kelumpuhan permanen yang memengaruhi kualitas hidup penderitanya secara drastis(Lishania et al., 2019).

Stroke bisa terjadi karena banyak faktor yang saling berkaitan. Salah satu penyebab utamanya adalah faktor usia, di mana sebagian besar terutama di Semarang penderita stroke berada di rentang usia 40 sampai 60 tahun (Lestari et al., 2020). Selain itu, mereka yang memiliki riwayat tekanan darah tinggi atau penyakit jantung juga memiliki risiko lebih besar. Gaya hidup juga sangat berpengaruh, seperti jarang bergerak atau berolahraga, kebiasaan merokok, dan pola hidup yang kurang sehat (Cenggono, 2025).

Di Indonesia sendiri, stroke menjadi salah satu penyakit yang banyak diderita oleh masyarakat. Bahkan menurut data dari WHO, stroke berada di peringkat ketiga sebagai penyakit paling mematikan setelah penyakit jantung dan kanker (Puspitawuri et al., 2019).

Banyak faktor yang dapat dimodifikasi telah terbukti memiliki pengaruh kumulatif terhadap risiko stroke dari 44% hingga 79%. Mayoritas orang dengan pendapatan rendah memilih untuk makan-makanan yang tidak sehat, seperti salmon asin, yang lebih murah tetapi memiliki kandungan garam yang tinggi, jika tertelan berlebihan, meningkatkan risiko stroke dan tekanan darah tinggi. Teori ini menunjukkan bahwa mengkonsumsi makanan ringan dan diet seimbang dapat membantu mencegah stroke (Susanti et al., 2024).

Berdasarkan teori kesehatan preventif dan perilaku konsumsi, pola makan dan gaya hidup memiliki kontribusi besar dalam meningkatkan atau menurunkan risiko penyakit tidak menular. Teori ini menunjukkan bahwa penerapan pola makan seimbang, berhenti merokok, serta gaya hidup aktif merupakan pendekatan strategis untuk mencegah stroke secara dini (Susanti et al., 2024).

Meskipun berbagai penelitian telah membahas prediksi risiko stroke, sebagian besar masih menggunakan pendekatan tradisional dengan variabel klinis terbatas. Selain itu, pendekatan teknologi seperti *machine learning* dalam prediksi stroke masih jarang diterapkan dalam konteks data lokal Indonesia, khususnya yang mempertimbangkan faktor sosial-ekonomi, gaya hidup, dan pola konsumsi. *Machine learning* merupakan salah satu cabang ilmu Kecerdasan Buatan (*Artificial Intelligence)* yang berkembang sangat cepat dan telah menyebabkan masalah klasifikasi, regresi, klastering, dan anomaly detection pada berbagai bidang dapat diatasi lebih efisien (Heryadi & Wahyono, 2020). Penelitian terdahulu menggunakan dataset penyakit stroke pernah dilakukan oleh Gangavarapu Sailasya dan Gorli L Aruna Kumari. Mereka melakukan tahap pengujian dengan menggunakan teknik *under sampling* serta membandingkan beberapa model yang diuji seperti KNN, *Decision Tree, Random Forest, SVM, Logistic Regression*, dan *Naive Bayes* dengan menghasilkan evaluasi model terbaik yakni menggunakan model Naive Bayes dengan akurasi sebesar 82% (Sailasya & Kumari, 2021). Hal ini menunjukkan adanya kesenjangan penelitian dalam hal pendekatan prediktif yang lebih komprehensif dan berbasis data lokal.

Penelitian ini bertujuan untuk membangun model prediksi penyakit stroke dengan menerapkan tiga algoritma klasifikasi berbasis *machine learning*, yaitu *Naïve Bayes*, *K-Nearest Neighbors* (KNN), dan *Decision Tree*. Ketiga algoritma ini dipilih karena masing-masing memiliki pendekatan berbeda dalam klasifikasi seperti K-Nearest Neighboors (KNN) yaitu metode klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut(Rizal et al., 2021), *Naïve Bayes* yaitu metode klasifikasi dengan menggunakan probabilitas dan statistic untuk memprediksi peluang di masa depan berdasarkan masa sebelumnya(Sani et al., 2022), dan *Decision Tree* merupakan struktur pohon yang terdiri dari node-node yang mempresentasikan keputusan dan

cabang-cabang yang mempresentasikan konsekuensi dari sebuah keputusan (Ramadhon et al., 2024). Selain itu, penelitian ini juga menerapkan teknik penyeimbangan data SMOTE-Tomek untuk menangani distribusi data yang tidak seimbang, serta mengevaluasi performa masing-masing algoritma guna menentukan model terbaik berdasarkan metrik evaluasi klasifikasi. Arah penelitian ini difokuskan pada optimalisasi proses klasifikasi pasien berisiko stroke berdasarkan faktor gaya hidup, demografi, dan riwayat kesehatan.

Kebaruan (*novelty*) dari penelitian ini terletak pada penerapan kombinasi teknik *resampling* dan perbandingan berbagai algoritma klasifikasi modern untuk memprediksi stroke pada data populasi Indonesia. Pendekatan ini belum banyak digunakan dalam studi serupa sebelumnya, sehingga diharapkan dapat memberikan kontribusi ilmiah dalam pengembangan sistem deteksi dini yang efisien dan berbiaya rendah.

Dengan mempertimbangkan urgensi kesehatan masyarakat, tingginya beban ekonomi, serta kesenjangan metodologis dalam studi terdahulu, penelitian ini diharapkan mampu memberikan alternatif pendekatan prediksi stroke yang lebih akurat, komprehensif, dan mudah diimplementasikan pada sistem layanan kesehatan berbasis digital. Bagian selanjutnya akan membahas metodologi yang digunakan untuk membangun dan mengevaluasi model prediktif tersebut.

METODE PENELITIAN

Metode penelitian merupakan gambaran umum mengenai alur atau langkah-langkah yang dilakukan dalam suatu penelitian (Tojiri et al., 2023). Penyusunan metode bertujuan agar pembaca dapat memahami setiap tahapan yang dilakukan oleh peneliti, mulai dari pengumpulan data, pengolahan data, pemodelan, hingga proses evaluasi. Selain itu, metode penelitian juga memberikan arahan dalam menentukan teknik analisis data yang tepat guna serta memperoleh hasil evaluasi yang akurat.

Penelitian ini dilakukan menggunakan pendekatan kuantitatif dengan metode *eksploratif-prediktif* berbasis *machine learning*. Data yang diambil dari penelitian ini merupakan data sekunder berasal dari repositori publik Kaggle dengan tautan: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. Dataset tersebut terdiri dari 5.110 baris data dengan total 12 kolom atau atribut, yaitu ID, umur, jenis kelamin, hipertensi, kondisi penyakit jantung, status pernikahan, status pekerjaan, tipe tempat tinggal, rata-rata kadar glukosa, BMI (*Body Mass Index*), status merokok, dan diagnosis penyakit stroke. Dari kedua belas atribut tersebut, sebanyak 10 atribut berperan sebagai variabel independen, sedangkan satu atribut yaitu "stroke" merupakan variabel dependen atau label. Kolom label ini memiliki dua nilai kelas, yaitu 0 dan 1, di mana angka 0 menunjukkan bahwa pasien tidak terdiagnosis stroke dan angka 1 menunjukkan bahwa pasien terdiagnosis stroke. Dataset ini menjadi dasar utama dalam membangun model prediksi pada penelitian yang dilakukan.

Populasi dalam penelitian ini merujuk pada seluruh data rekam medis pasien yang memiliki atribut atau karakteristik yang serupa dengan yang tercantum dalam dataset, yaitu atribut yang berkaitan dengan kondisi kesehatan dan faktor risiko stroke. Karena penelitian ini bersifat sekunder dan menggunakan data yang sudah tersedia di repositori *kaggle*, maka sampel yang digunakan adalah keseluruhan data yang tersedia dalam dataset, yaitu sebanyak 5.110 entri data. Dengan demikian, teknik sampling yang digunakan adalah sampling jenuh (total sampling), di mana semua anggota populasi dijadikan sampel.

Teknik pengumpulan data yang digunakan dalam penelitian ini adalah dokumentasi dan pengambilan data sekunder daring. Data diperoleh secara langsung dari situs Kaggle yang merupakan repositori dataset yang dapat diakses oleh publik. Peneliti tidak melakukan eksperimen atau wawancara langsung terhadap subjek penelitian, melainkan menggunakan data yang telah tersedia dan dipublikasikan oleh penyedia dataset.

Langkah awal dalam penelitian ini, yakni *Pre-processing* dengan mengolah data mentah agar dapat digunakan dalam pelatihan model (Perwitasari et al., 2023). Langkah pertama dalam *preprocessing* yakni *replace missing value*, yaitu mengisi data yang masih kosong dengan nilai median dari atribut terkait. Langkah selanjutnya adalah *replace value*, yang bertujuan untuk mengganti nilai tidak diketahui atau "*unknown*" menjadi nilai yang valid. Dalam penelitian ini, nilai "*unknown*" pada atribut smoking_status diubah menjadi "*formerly smoked*", sedangkan pada atribut gender, nilai "*other*" diganti menjadi "*female*".

Tahap berikutnya adalah *feature engineering*, yaitu proses pengelompokan nilai pada atribut umur menjadi lima kategori berdasarkan rentang usia, yaitu *Teenager, Adult, Middle Age, Elderly*, dan

Old. Setelah itu, dilakukan proses *label encoding*, di mana seluruh atribut yang berbentuk kategori (string) dikonversi ke dalam bentuk bilangan bulat (integer). Konversi ini diperlukan agar data dapat diproses oleh algoritma pembelajaran mesin (*machine learning*). Selanjutnya dilakukan *feature correlation*, yakni proses untuk menganalisis tingkat korelasi antara setiap atribut dengan label. Atribut yang memiliki korelasi paling rendah terhadap label akan dihapus agar performa model menjadi lebih stabil dan optimal.

Setelah data selesai melalui proses *data preprocessing*, langkah berikutnya adalah mengatasi ketidakseimbangan data (*imbalanced dat*a). Dalam hal ini, data menunjukkan ketimpangan jumlah antara kelas label 0 dan 1, di mana jumlah data pasien yang tidak terkena stroke jauh lebih banyak daripada yang terkena stroke. Untuk mengatasi hal ini, digunakan teknik *oversampling* dengan metode SMOTE-TOMEK (*Synthetic Minority Oversampling Technique – Tomek Links*). SMOTE bertujuan untuk memperbanyak data pada kelas minoritas secara sintetis agar distribusi data menjadi lebih seimbang sehingga model dapat belajar dengan baik tanpa bias terhadap salah satu kelas(Kamalov et al., 2025). Sedangkan, TOMEK bertujuan untuk teknik *Tomek Links* digunakan untuk membersihkan *noise* dan *overlap* antar kelas. Data pasangan yang terlalu dekat dihapus untuk memperjelas batas antar kelas

Tahap selanjutnya adalah pembagian data ke dalam dua bagian, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Pembagian ini bertujuan untuk mengukur performa model pada data yang belum pernah dilihat sebelumnya, serta mendeteksi potensi *overfitting* atau *underfitting*. Dalam penelitian ini, proporsi data dibagi sebesar 80% untuk pelatihan dan 20% untuk pengujian. Model akan dilatih menggunakan data *training* dan kemudian diuji menggunakan data *testing* untuk melihat kinerjanya secara objektif (Lin et al., 2017).

Data yang telah melalui tahap *preprocessing* dan penanganan *imbalance data* menggunakan SMOTE-TOMEK, kemudian digunakan dalam pelatihan beberapa algoritma pembelajaran mesin. Adapun algoritma yang diterapkan dalam penelitian ini meliputi *Decision Tree*, *Naive Bayes*, dan K-*Nearest Neighbors* (k-NN). Pemilihan berbagai algoritma ini bertujuan untuk membandingkan performa masing-masing model dalam memprediksi penyakit stroke berdasarkan dataset yang tersedia.

Evaluasi model dilakukan untuk mengetahui sejauh mana model yang dibangun dapat memprediksi stroke secara akurat. Pengukuran performa dilakukan menggunakan beberapa metrik evaluasi, yaitu Accuracy, Precision, Recall, dan F1-Score. Metrik-metrik tersebut memberikan gambaran menyeluruh mengenai kemampuan klasifikasi model terhadap data yang telah disediakan. Dengan demikian, dapat diketahui algoritma mana yang memberikan hasil terbaik untuk digunakan dalam prediksi penyakit stroke.

HASIL DAN PEMBAHASAN

Hasil

Dalam penelitian ini mengambil dari dataset penyakit stroke yang tersedia di *kaggle* dengan berisikan informasi pasien yang diprediksi memiliki riwayat penyakit stroke. Data ini diambil dari data publik dengan laman web https://www.kaggle.com/fedesoriano/stroke-prediction-dataset dengan jumlah atribut 12 dan 5.110 data.

Tabel 1 Informasi Dataset Stroke

Nama Dataset	Jumlah Atribut	Jumlah Record
Stroke Dataset Prediction	12	5.110

Sumber: *Kaggle.com* https://www.kaggle.com/fedesoriano/stroke-prediction-dataset)

Adapun informasi lengkap dari setiap atribut yang terdapat pada dataset Stroke *Prediction.*

Tabel 2 Informasi Atribut Pada Dataset Stroke

No	Nama Atribut	Tipe Data	Keterangan
1	*Id*	*int64*	Nomor Pengenal
2	*gender*	*Object*	Jenis Kelamin Pasien
3	*Age*	*float64*	Usia Pasien
4	*hypertension*	*int64*	Riwayat Hipertensi
5	*heart_disease*	*int64*	Riwayat Penyakit Jantung

No	Nama Atribut	Tipe Data	Keterangan
6	*ever_married*	*Object*	Status Pernikahan
7	*work_type*	*Object*	Tipe Pekerjaan
8	*residence_type*	*Object*	Kondisi Tempat Tinggal
9	*avg_glucose_level*	*float64*	Rata-rata level glukosa
10	*Bmi*	*float64*	Berat Badan
11	*smoking_status*	*Object*	Status Merokok
12	*stroke*	*int64*	Status Stroke

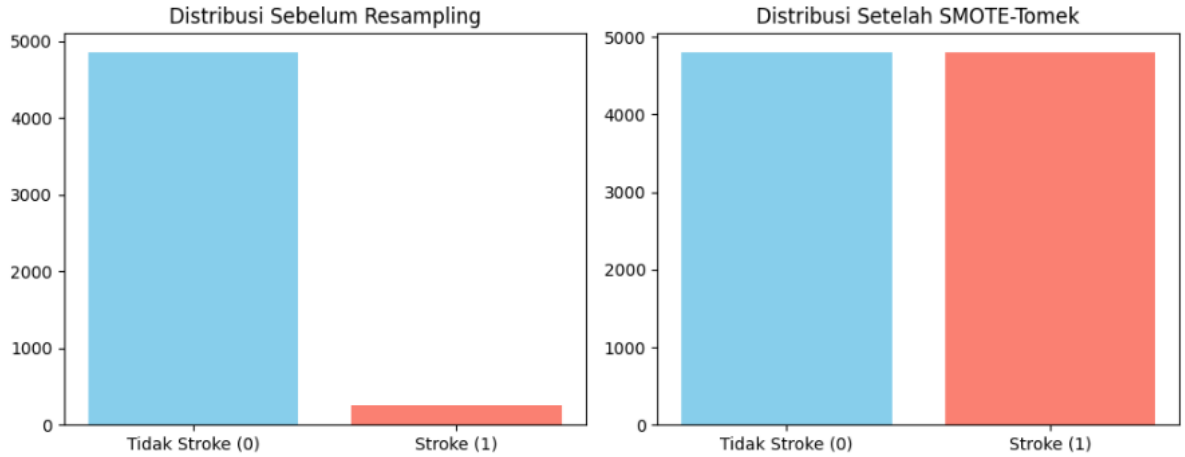
Sumber: Kaggle.com https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

1. *Pre-processing data*

Dalam tahap *pre-processing data* menjadi salah satu tahapan awal pada *data mining* untuk mengubah data mentah menjadi informasi yang lebih bersih dan dapat digunakan untuk pengolahan selanjutnya. Dalam penelitian ini, peneliti melakukan *pre-processing* data dengan beberapa tahapan, diantaranya melakukan *Replace Missing Value, Replace Value, Feature Engineering, Feature Correllation* dan *label Encoder.*

2. Teknik *Handling Imbalance Data*

Setelah *preprocessing* data selesai, maka tahapan selanjutnya yakni menangani data yang memiliki kelas *imbalance* dengan cara menggunakan teknik *oversampling*. Dimana pada penelitian ini menggunakan teknik oversampling yaitu *SMOTE-TOMEK.*



Gambar 1 Perbandingan Sebelum dan Sesudah *Handling Imbalance Data*
Sumber: Hasil Olahan Data (2025)

Setelah proses *preprocessing* selesai, langkah selanjutnya adalah menangani permasalahan ketidakseimbangan kelas pada data. Teknik *oversampling* yang digunakan dalam penelitian ini adalah SMOTE-Tomek, yang menggabungkan *Synthetic Minority Oversampling Technique* (SMOTE) dengan metode *Tomek Links* untuk menyeimbangkan distribusi data antar kelas. Efektivitas teknik ini dapat dilihat pada Gambar 1 yang menunjukkan perbandingan distribusi kelas sebelum dan sesudah dilakukan penanganan data *imbalance*. Terlihat bahwa jumlah data pada kelas minoritas meningkat secara signifikan setelah diterapkannya metode tersebut, sehingga diharapkan dapat meningkatkan kinerja model klasifikasi.

3. Model Selection (*Test_Train Split*)

Kemudian jika kelas sudah *balance*, maka tahapan berikutnya yakni *Model Selection*, dimana data dibagi menjadi *training* dan *testing*. Hal ini digunakan untuk melihat hasil pada data *training* dan *testing* apakah mengalami *overfitting* atau *underfitting* pada algoritma yang akan digunakan. Pada penelitian ini dilakukan pembagian data *training* yaitu sebanyak 80% dari data dan *testing* sebanyak 20%.

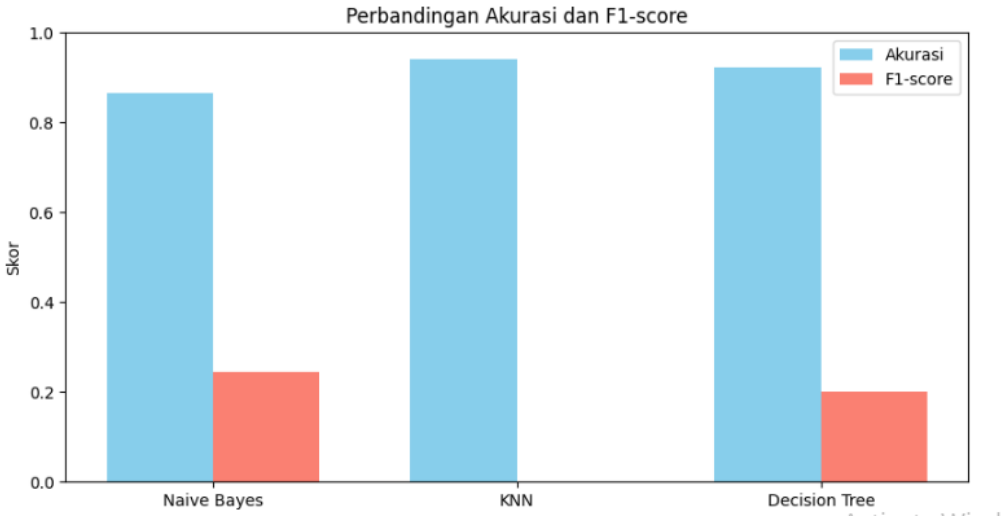
Data yang sudah melalui proses *model selection* (*test train split*) akan dilatih dengan model. Penelitian ini menggunakan beberapa model yang dilatih. Adapun model yang dipakai dalam pelatihan antara lain *Decision Tree, K-Nearest Neighbors* (KNN), dan *Naïve Bayes*. Ketiga algoritma ini dipilih karena memiliki pendekatan yang berbeda dalam melakukan klasifikasi, sehingga diharapkan dapat memberikan gambaran yang beragam terhadap performa prediksi pada data yang telah seimbang.

Decision Tree mampu menangani data dengan fitur kategorik maupun numerik serta memberikan interprestasi yang jelas dalam bentuk pohon keputusan. KNN bekerja berdasarkan kedekatan antar data, sehingga performanya sangat dipengaruhi oleh representasi data yang telah dibersihkan dan seimbang. Sementara itu, *Naïve Bayes* yang berbasis probabilistik sangat efektif pada data dengan dimensi tinggi dan distribusi yang cukup baik setelah *oversampling.* Dengan membandingkan ketiganya, penelitian ini bertujuan untuk mengetahui model mana yang paling sesuai dengan karakteristik data yang telah diproses, khusunya dalam konteks prediksi penyakit stroke.

4. Evaluasi Model

a. Sebelum melakukan Teknik *Imbalance data* SMOTE-TOMEK

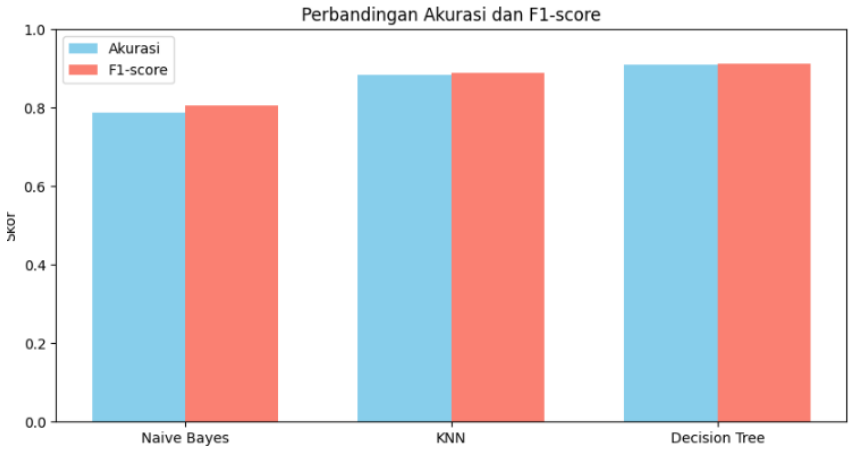
Sebelum diterapkannya teknik SMOTE-TOMEK, distribusi data pada kelas target sangat timpang. Mayoritas data tergolong dalam kelas non-stroke (label 0), sementara data dengan label stroke (label 1) jumlahnya jauh lebih sedikit. Ketidakseimbangan ini menyebabkan beberapa model cenderung bias terhadap kelas mayoritas dan menghasilkan metrik evaluasi seperti *recal*l dan F1-*score* yang sangat rendah atau bahkan nol. Hal ini terlihat jelas dari hasil evaluasi awal yang menunjukkan akurasi tinggi namun tidak diimbangi dengan nilai recall maupun F1-score yang memadai.



Gambar 2. Perbandingan Akurasi dan F1-*Score* sebelum menggunakan SMOTE-TOMEK
Sumber: Hasil Olahan Data (2025)

b. Setelah melakukan Teknik *Imbalance* data SMOTE-TOMEK

Setelah dilakukan penanganan data imbalance dengan teknik SMOTE-TOMEK, distribusi antara kelas 0 dan 1 menjadi lebih seimbang. Teknik ini tidak hanya menambah data pada kelas minoritas (SMOTE), tetapi juga membersihkan pasangan data yang tumpang tindih antar kelas (*Tomek Links*), sehingga menghasilkan dataset yang lebih bersih dan seimbang. Dampaknya terlihat pada peningkatan performa model secara signifikan, terutama pada nilai recall dan F1-*score*. Model seperti *Decision Tree* menunjukkan peningkatan yang paling menonjol, dan akurasi model menjadi lebih representatif terhadap performa sebenarnya.



Gambar 3. Perbandingan Akurasi dan F1-*Score* setelah menggunakan SMOTE-TOMEK
Sumber: Hasil Olahan Data (2025)

5. *Evaluation Model*

Pada tahap ini, dilakukan evaluasi pada model yang dibuat. Ada tiga model yang dilatih untuk mendapatkan hasil klasifikasi yang diinginkan antara lain *Decision Tree*, *Naïve Bayes*, dan k-NN, serta melakukan perbandingan dari teknik *imbalance* data dengan tidak menggunakan teknik *imbalance*.

Tabel 3 Evaluasi Model

Algoritma	Tidak Menggunakan SMOTE-Tomek		Menggunakan SMOTE-Tomek	
	Accuracy	F-1 Score	Accuracy	F-1 Score
Decision Tree	92%	20%	91%	91%
K-Nearest Neighbour	94%	0%	89%	88%
Naive Bayes	87%	24%	79%	81%

Sumber: Hasil Olahan Data (2025)

Penerapan SMOTE-Tomek menunjukkan dampak signifikan terhadap keseimbangan performa model klasifikasi. Tanpa SMOTE-Tomek, ketiga algoritma memang menunjukkan akurasi tinggi, namun F1-score sangat rendah, terutama pada K-Nearest Neighbour hanya 0% dan *Decision Tree* 20%, menandakan model bias terhadap kelas mayoritas. Setelah penerapan SMOTE-Tomek, meskipun akurasi sedikit menurun, *F1-score* meningkat drastis pada semua algoritma, seperti pada *Decision Tree* menjadi 91% dan KNN menjadi 88%. Ini menunjukkan bahwa SMOTE-Tomek berhasil memperbaiki ketidakseimbangan data, sehingga model menjadi lebih adil dan mampu mengenali kedua kelas dengan lebih baik.

6. Perbandingan Studi Terdahulu

Salah satu jurnal yang membahas tentang hal yang sama dengan menggunakan beberapa model klasifikasi, terlihat dari table berikut (Sailasya & Kumari, 2021).

Tabel 4. Perbandingan Studi Terdahulu

Algoritma	Penelitian oleh Gangavarapu dan Gorli	Penelitian yang dilakukan
Decision Tree	66%	91%
K-Nearest Neighbour	80%	89%
Naive Bayes	82%	79%

Sumber: Hasil Olahan Data (2025)

Maka dari itu, alasan memilih *decision tree* memiliki nilai akurasi yang paling unggul karena hasil dari penilitian ini dan setelah membandingkan dengan penelitian terdahulu mendapatkan nilai akurasi yang tinggi yakni 91%.

Pembahasan

Berdasarkan hasil evaluasi model sebelum dan sesudah dilakukan penanganan data tidak seimbang menggunakan teknik SMOTE-Tomek, terlihat adanya peningkatan yang cukup signifikan khususnya pada model *Decision Tree*. Nilai *F1-Score* meningkat dari 20% menjadi 91% setelah teknik SMOTE-Tomek diterapkan. Hal ini menunjukkan bahwa teknik *balancing data* yang digunakan berhasil memperbaiki distribusi data pada kelas minoritas (stroke), sehingga model mampu mengenali pasien yang memiliki risiko stroke dengan lebih baik.

Peningkatan ini menjadi hal yang penting dalam konteks prediksi medis, karena kesalahan dalam mengklasifikasikan pasien stroke (khususnya jika model gagal mengenali pasien stroke atau *false negative*) dapat menimbulkan risiko yang serius. Oleh karena itu, peningkatan nilai *recall* dan *F1-score* pada model setelah proses *balancing* menunjukkan bahwa teknik SMOTE-Tomek mampu meningkatkan keakuratan sistem prediksi stroke secara keseluruhan.

Jika dibandingkan dengan penelitian yang dilakukan sebelumnya, Algoritma Naive Bayes memperoleh hasil terbaik dengan akurasi sebesar 82% (Sailasya & Kumari, 2021). Namun, dalam penelitian ini, model *Decision Tree* justru memberikan hasil paling optimal dengan akurasi 91% dan *F1-Score* yang jauh lebih tinggi. Perbedaan hasil ini diduga disebabkan oleh penggunaan metode *balancing* yang berbeda, di mana penelitian ini menggunakan teknik SMOTE-Tomek, sedangkan penelitian sebelumnya menggunakan teknik undersampling. Selain itu, penelitian ini juga melakukan tahapan preprocessing yang lebih lengkap, seperti pengelompokan usia *(feature engineering*), proses *encoding* data kategorikal, dan seleksi fitur berdasarkan korelasi.

Temuan ini diperkuat oleh teori yang disampaikan oleh Kamalov, Choutri, dan Atiya (2025) yang menjelaskan bahwa teknik SMOTE-Tomek tidak hanya efektif dalam menyeimbangkan data, tetapi juga membantu membersihkan noise dan data yang tumpang tindih antar kelas, sehingga hasil pelatihan model menjadi lebih akurat.

Secara umum, hasil penelitian ini memberikan implikasi bahwa dalam membangun sistem prediksi penyakit dengan karakteristik data yang tidak seimbang, penggunaan teknik *balancing data* seperti SMOTE-Tomek sangat direkomendasikan. Selain itu, algoritma *Decision Tree* terbukti memberikan hasil yang cukup tinggi dan dapat digunakan sebagai model yang mudah diinterpretasikan, khususnya dalam implementasi sistem deteksi dini stroke di dunia nyata.

KESIMPULAN

Penelitian ini menunjukkan bahwa penggunaan teknik penanganan data tidak seimbang menggunakan SMOTE-TOMEK (*Tomek* (*Synthetic Minority Over-sampling Technique - Tomek Links*) mampu meningkatkan performa prediksi penyakit stroke secara signifikan. Proses ini dilakukan dengan membandingkan enam algoritma klasifikasi, dan *Decision Tree* menjadi model terbaik dengan hasil akurasi 91% menjadi evaluasi paling optimal.

Secara keseluruhan, SMOTE-TOMEK terbukti mampu mengatasi masalah distribusi kelas yang timpang, meningkatkan sensitivitas model terhadap kasus stroke, dan memberikan hasil evaluasi yang lebih stabil dibandingkan tanpa *handling imbalance data* atau pendekatan *under sampling*. Penelitian ini berkontribusi dalam memberikan pendekatan yang lebih akurat dalam sistem deteksi dini penyakit stroke menggunakan metode data mining, khususnya untuk klasifikasi berbasis data tidak seimbang.

DAFTAR PUSTAKA

Anisah Makkiyah, F. (2024). Kenali stroke, gejala, serta pencegahan stroke. *SEGARA: Jurnal Pengabdian Kepada Masyarakat*, *1*(2), 53–55. https://doi.org/10.33533/segara.v1i2.7260

Cenggono, M. (2025). *Pengaruh Faktor Risiko dan Gaya Hidup terhadap Risiko Stroke*. *9*, 6218–6227.

Chavda, V., Chaurasia, B., Deora, H., & Umana, G. E. (2021). Chronic Kidney disease and stroke: A Bi-directional risk cascade and therapeutic update. *Brain Disorders*, *3*(March), 100017. https://doi.org/10.1016/j.dscb.2021.100017

Heryadi, Y., & Wahyono, T. (2020). *Machine Learning: Konsep dan Implementasi*. *August*.

Kamalov, F., Choutri, S. E., & Atiya, A. F. (2025). Analytical Formulation of Synthetic Minority Oversampling Technique (Smote) for Imbalanced Learning. *Gulf Journal of Mathematics*, *19*(1),

400–415. https://doi.org/10.56947/gjom.v19i1.2639

Lestari, L. M., Pudjonarko, D., & Handayani, F. H. (2020). Characteristics of stroke patients: An analytical description of outpatient at the hospital in Semarang Indonesia. *Jurnal Aisyah : Jurnal Ilmu Kesehatan*, *5*(1), 67–74. https://doi.org/10.30604/jika.v0i0.287

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, *5*, 16568–16575. https://doi.org/10.1109/ACCESS.2017.2738069

Lishania, I., Goejantoro, R., & Nasution, Y. N. (2019). Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda. *Jurnal Eksponensial*, *10*(2), 135–142.

Perwitasari, A., Septiriana, R., & Tursina, T. (2023). Data preparation Structure untuk Pemodelan Prediktif Jumlah Peserta Ajar Matakuliah. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *9*(1), 7. https://doi.org/10.26418/jp.v8i3.57321

Puspitawuri, A., Santoso, E., & Dewi, C. (2019). Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(4), 3319–3324.

Ramadhon, R. N., Ogi, A., Agung, A. P., Putra, R., Febrihartina, S. S., & Firdaus, U. (2024). Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank. *Karimah Tauhid*, *3*(2), 1860–1874. https://doi.org/10.30997/karimahtauhid.v3i2.11952

Riyadina, W., Pradono, J., Kristanti, D., & Turana, Y. (2020). Stroke in Indonesia: Risk factors and predispositions in young adults. *J Cardiovasc Dis Res*, *2*(11), 178–183.

Rizal, Aidilof, H. A. K., & Kurniawan, W. (2021). Klasifikasi Berita Olahraga Pada Portal Berita Online Dengan Metode K-Nearest Neighbour (KNN) Dan Levenshtein Distance. *Jurnal Teknologi Terapan and Sains 4.0*, *2*(1), 365. https://doi.org/10.29103/tts.v2i1.3760

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 539–545. https://doi.org/10.14569/IJACSA.2021.0120662

Sani, Hafara, A., Setiawan, A., & Primadewi, A. (2022). Penerapan Metode Naive Bayes Dalam Rekomendasi Strategi Penerimaan Peserta Didik Baru. *Journal of Computer System and Informatics (JoSYC)*.

Susanti, N., Vinanda, F., Andini, & Syahfitri, W. (2024). Pengaruh Gaya Hidup Sehat Terhadap Pencegahan Stroke. *Ibnu Sina: Jurnal Kedokteran Dan Kesehatan-Fakultas Kedokteran Universitas Islam Sumatera Utara*, *23*(2), 20353.

Tojiri, Y., Putra, H. S., & Faliza, N. (2023). Buku Dasar Metodologi Penelitian: Teori Desain dan Analisis Data. In *Takaza Innovatix Labs* (Issue January).

Turana, Y., Tengkawan, J., Chia, Y. C., Nathaniel, M., Wang, J., Sukonthasarn, A., Chen, C., Minh, H. Van, Buranakitjaroen, P., Shin, J., Siddique, S., Nailes, J. M., Park, S., Teo, B. W., Sison, J., Ann Soenarta, A., Hoshide, S., Tay, J. C., Prasad Sogunuru, G., … Kario, K. (2021). Hypertension and stroke in Asia: A comprehensive review from HOPE Asia. *The Journal of Clinical Hypertension*, *23*(3), 513–521. https://doi.org/10.1111/jch.14099