

Klasifikasi Penyakit Stroke Menggunakan Algoritma XGBoost Classifier

Ani Dijah Rahajoe¹, Fairus Irhab Adinata E fendi², Azmi Maulana Mahardika³, Muhammad Rizky Firdaus⁴, Achmad Ario Dwi Maulana⁵, Piter Rudi Irson Rivaldo Lapon⁶

^{1,2,3,4,5,6}Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur,

Corresponding Author: Muhammad Rizky Firdaus, e-mail: 24081010077@student.upnjatim.ac.id

Published: January, 2026

ABSTRAK

Stroke merupakan krisis kesehatan global yang menjadi penyebab utama kematian dan kecacatan. Deteksi lebih dini faktor stroke sangat penting untuk tindakan pencegahan yang efektif. Studi ini bertujuan untuk menganalisis karakteristik dataset stroke, mengkaji pengaruh teknik Hybrid Sampling (gabungan SMOTE dan Random UnderSampling) dalam mengatasi ketidakseimbangan data, serta mengevaluasi performa mode XGBoost. Penelitian ini menggunakan pendekatan kuantitatif eksperimen dengan dataset sekunder dari Kaggle. Penanganan ketidak seimbangan kelas dilakukan dengan Hybrid Sampling (gabungan SMOTE dan Random UnderSampling). Penerapan teknik Hybrid Sampling ini terbukti memberikan dampak signifikan pada kemampuan deteksi model, di mana nilai F1-Score pada kelas minoritas (stroke) yang semula bernilai 0,00 pada data asli (tanpa penanganan ketidakseimbangan) berhasil meningkat menjadi 0,25. Model dikembangkan menggunakan algoritma Extreme Gradient Boosting (XGBoost). Dataset memiliki ketidakseimbangan ekstrem dengan 4,8% pada kelas strok. Hasil evaluasi menunjukkan model XGBoost mencapai akurasi sebesar 89,4% pada data uji. Namun, performa pada kelas minoritas (stroke) masih rendah dengan recall 0,36 dan F1-Score 0,25. Meskipun XGBoost memberikan akurasi tinggi secara keseluruhan data yang signifikan. Diperlukan optimasi lebih lanjut melalui hyperparameter tuning atau algoritma boosting lainnya.

Keywords: Prediksi Risiko Stroke, Algoritma XGBoost Classifier, Hybrid Sampling, Klasifikasi.

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

PENDAHULUAN

Stroke merupakan salah satu krisis kesehatan global yang menduduki peringkat tinggi sebagai penyebab utama kematian dan kecacatan di seluruh dunia (World Health Organization, 2023; Kementerian Kesehatan Republik Indonesia, 2019). Penyakit ini terjadi akibat gangguan aliran darah ke otak yang secara progresif dapat menyebabkan kerusakan jaringan permanen apabila tidak diidentifikasi dan ditangani secara cepat (World Health Organization, 2023). Faktor risiko stroke dapat berasal dari kondisi medis seperti hipertensi dan diabetes mellitus, serta faktor gaya hidup seperti pola makan tidak sehat, kurangnya aktivitas fisik, dan merokok (Kianie & Cenggono, 2025).

Laporan berbagai organisasi kesehatan dunia menegaskan bahwa stroke tetap menjadi tantangan medis utama yang memicu kecacatan jangka panjang, baik di negara maju maupun berkembang, termasuk Indonesia (World Health Organization, 2023; Kementerian Kesehatan Republik Indonesia, 2019). Mengingat dampaknya yang luas, upaya pencegahan dan penanganan dini sangat bergantung pada kemampuan sistem kesehatan dalam melakukan deteksi terhadap faktor-faktor risiko yang dimiliki setiap individu secara akurat (Kementerian Kesehatan Republik Indonesia, 2019).

Namun, proses diagnosis konvensional seringkali menghadapi kendala efisiensi karena memerlukan pemeriksaan klinis lanjutan dan pencitraan medis yang memakan waktu, biaya, serta sumber daya tenaga medis yang signifikan (Kementerian Kesehatan Republik Indonesia, 2019). Kesenjangan ini menjadi tantangan serius, terutama pada fasilitas pelayanan kesehatan dengan keterbatasan infrastruktur dan tenaga ahli (Kementerian Kesehatan Republik Indonesia, 2019). Di sisi lain, kemajuan pesat dalam teknologi informasi, khususnya dalam disiplin pembelajaran mesin (*machine learning*), telah membuka peluang baru untuk memanfaatkan data medis dalam membantu proses pengambilan keputusan klinis (Rajkomar, Dean, & Kohane, 2019). Model pembelajaran mesin terbukti mampu mempelajari pola kompleks dari data historis pasien untuk menghasilkan prediksi risiko penyakit secara lebih cepat, objektif, dan terukur (Rajkomar, Dean, & Kohane, 2019). Meskipun demikian, pemodelan data stroke menghadapi tantangan metodologis berupa ketidakseimbangan distribusi kelas, di mana jumlah data pasien non-stroke jauh lebih dominan dibandingkan dengan data pasien stroke (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; He & Garcia, 2009).

Oleh karena itu, diperlukan sebuah pendekatan pemodelan yang tidak hanya memiliki performa tinggi, tetapi juga mampu menangani anomali data tersebut. Algoritma *Extreme Gradient Boosting* (XGBoost) hadir sebagai salah satu metode *ensemble learning* yang dikenal sangat efektif dalam menangani data tabular kompleks dan menghasilkan kinerja prediktif yang superior melalui teknik regularisasi (Chen & Guestrin, 2016). Beberapa penelitian sebelumnya telah membahas prediksi stroke menggunakan berbagai metode *machine learning*. Misalnya, penelitian berjudul “Prediksi Stroke Menggunakan Deep Neural Network dan Oversampling SMOTE” menunjukkan bahwa DNN efektif menangani ketidakseimbangan data (Faisal & Subekti, 2022), sedangkan studi “Klasifikasi Risiko Stroke dengan Support Vector Machine” menemukan SVM mampu memberikan akurasi tinggi (Nabila & Pamungkas, 2023). Selain itu, beberapa penelitian lain membandingkan algoritma seperti Random Forest dan XGBoost, yang terbukti unggul dalam menangani fitur kompleks dan dataset tidak seimbang (Dritsas & Trigka, 2022). Untuk memperkuat akurasi pada kelas minoritas, penerapan teknik Hybrid Sampling (gabungan SMOTE dan Random UnderSampling) pada tahap pra-pemrosesan data menjadi krusial untuk meningkatkan representasi data pasien stroke sehingga proses pembelajaran model menjadi lebih adil dan seimbang (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Berdasarkan urgensi tersebut, penelitian ini bertujuan untuk menganalisis karakteristik dataset stroke serta mengkaji pengaruh integrasi teknik Hybrid Sampling (gabungan SMOTE dan Random UnderSampling) dalam menetralkan bias data. Secara khusus, studi ini mengevaluasi sejauh mana algoritma XGBoost dapat dioptimalkan untuk memprediksi risiko stroke dengan fokus pada metrik yang sensitif terhadap kelas minoritas, seperti AUC, *precision*, *recall*, dan F1-score. Melalui pendekatan ini, diharapkan dapat dihasilkan sebuah kerangka kerja prediktif yang andal sebagai instrumen deteksi dini stroke guna menekan angka kematian dan disabilitas akibat penyakit ini.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimen dengan fokus pada penerapan pembelajaran mesin untuk memprediksi risiko stroke berdasarkan data medis pasien. Data yang digunakan adalah data sekunder yang mencakup informasi faktor risiko stroke seperti usia, tekanan darah, kadar kolesterol, riwayat keluarga, dan faktor-faktor klinis lainnya. Dalam penelitian ini, diterapkan teknik Hybrid Sampling (gabungan SMOTE dan Random UnderSampling) untuk mengatasi masalah ketidakseimbangan distribusi kelas antara kasus stroke dan non-stroke. Hybrid Sampling bertujuan untuk menghasilkan data sintetis pada kelas minoritas, sehingga model pembelajaran mesin dapat dilatih dengan dataset yang lebih seimbang, meningkatkan kemampuan model dalam mengklasifikasikan kasus positif (stroke) dengan lebih akurat.

Model prediksi dibangun menggunakan algoritma XGBoost (Extreme Gradient Boosting), yang terkenal efisien dalam menangani data yang kompleks dan menghasilkan prediksi yang akurat. Proses penelitian mencakup tahapan pra-pemrosesan data, yang meliputi pembersihan data, pengisian nilai yang hilang, dan standarisasi fitur. Setelah model dilatih menggunakan data pelatihan, model dievaluasi dengan metrik AUC, precision, recall, dan F1-score untuk menilai performa dalam memprediksi risiko stroke. Selanjutnya, hasil model XGBoost dibandingkan dengan model klasifikasi konvensional seperti regresi logistik atau decision tree untuk melihat sejauh mana penerapan XGBoost dapat meningkatkan akurasi prediksi risiko stroke.

HASIL DAN PEMBAHASAN

Data Preprocessing

Data yang digunakan dalam penelitian ini diperoleh dari platform kaggle dengan judul Healthcare Dataset Stroke Data. Dataset ini dikategorikan sebagai data sekunder dan telah melalui proses anonimisasi untuk menjaga kerahasiaan identitas pasien. Dataset memiliki total 5110 baris data dan 12 kolom atribut, yang terdiri dari 11 fitur prediktor (variabel independen) dan 1 variabel target (dependen). Proses penanganan missing value dilakukan pada atribut BMI yang memiliki 201 data kosong menggunakan nilai rata-rata (mean)

Table 1. Attribute

	Tipe Data	Deskripsi
Id	Numerik Diskrit	Pengenal unik (diabaikan dalam pemodelan)
Gender	Kategorikal	Jenis kelamin (Male, Female, Other)
Age	Numerik Kontinu	Usia pasien dalam satuan tahun
Hypertension	Biner (0/1)	1 jika pasien menderita hipertensi, 0 jika tidak
Heart_disease	Biner (0/1)	1 jika pasien menderita penyakit jantung, 0 jika tidak
Ever_married	Kategorikal	Status pernikahan subjek (Yes/ No)
Work_type	Kategorikal	Jenis pekerjaan (Govt_job, Private, Self-employed, dll)
Residence_type	Kategorikal	Tipe lokasi tempat tinggal (Urban/Rural)
Avg_glucose_level	Numerik Kontinu	Rata-rata kadar glukosa dalam darah
Bmi	Numerik Kontinu	Indeks Massa Tubuh (Body Mass Index)
Smoking_status	Kategorikal	Riwayat merokok (never smoked, formerly smoked, smokes, unknown)
Stroke	Biner (0/1)	Variabel target: 1 jika mengalami stroke, 0 jika tidak

Tahap pra-pemrosesan data dimulai dengan eliminasi fitur yang tidak relevan, yaitu atribut id karena tidak memberikan informasi signifikan bagi model prediksi. Selanjutnya, dilakukan penanganan data hilang (missing value) pada atribut bmi yang berjumlah 201 entri dengan menerapkan metode imputasi nilai rata-rata (mean). Untuk mempersiapkan data kategorikal ke dalam format numerik yang dapat diproses oleh algoritma XGBoost, dilakukan teknik Label Encoding pada variabel gender, ever_married, work_type, residence_type, dan smoking_status. Keseluruhan proses ini menghasilkan

dataset yang terstandarisasi dan siap untuk tahap analisis korelasi serta pemodelan lebih lanjut.

Exploratory Data Analysis (EDA)

a. Distribusi Data

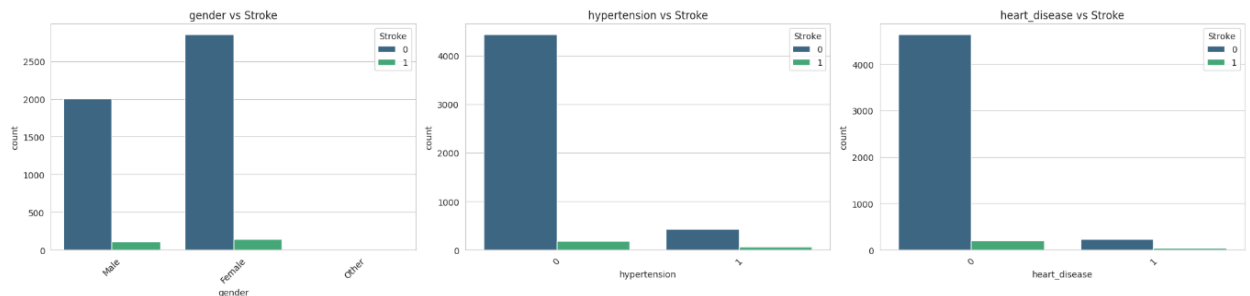


Figure 1. Hubungan Gender, Hypertension dan Penyakit Jantung Terhadap Risiko Kejadian Stroke.

Analisis terhadap faktor menunjukkan pasien perempuan mendominasi dataset secara kuantitas, rasio kejadian stroke antara laki-laki dan laki-laki tampak serupa, sehingga gender tidak menjadi faktor pembeda tunggal yang kuat tanpa kombinasi fitur lain. Sebaliknya, variabel medis seperti hipertensi dan penyakit jantung menunjukkan korelasi yang jauh lebih signifikan, meskipun mayoritas subjek tidak memiliki riwayat kedua penyakit tersebut, kelompok yang menderita hipertensi dan gangguan kardiovaskular menunjukkan rasio kejadian stroke yang jauh lebih tinggi secara proporsional. Hal ini mengonfirmasi bahwa tekanan darah tinggi dan kondisi kesehatan jantung merupakan indikator risiko klinis utama yang secara drastis meningkatkan peluang terjadinya serangan stroke dibandingkan faktor demografi jenis kelamin.

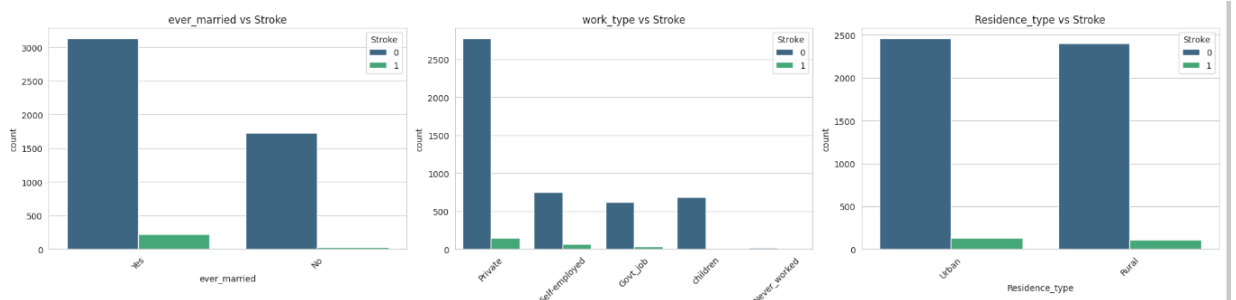


Figure 2. Hubungan Status Pernikahan, Jenis Pekerjaan, dan Tempat Tinggal Terhadap Risiko Kejadian Stroke.

Berdasarkan hasil Visualisasi Hubungan Antara Variabel Kategorik dengan Risiko Kejadian Stroke, status pernikahan dan jenis pekerjaan sejalan dengan anggapan awal bahwa faktor ini memiliki hubungan erat dengan risiko terjadinya stroke, yang dipengaruhi oleh faktor usia dan gaya hidup. Hal tersebut terlihat dari kedua variabel tersebut, individu yang pernah menikah dan karyawan swasta memiliki persentase terbesar kejadian stroke dikarenakan mereka dominan dengan rentang usia dewasa dan lansia. Sementara variabel ketiga menunjukkan sebaliknya yakni kedua kelompok yaitu anak-anak dan yang belum bekerja tidak ada yang mengalami kejadian stroke, teori dimana risiko terjadi stroke berkaitan dengan usia dan tekanan internal dalam bekerja mempengaruhi kejadian stroke. Terakhir, variabel tempat tinggal tidak ada yang signifikan sebab distribusi kejadian stroke pada status tempat tinggal memiliki rentang nilai yang relatif seimbang antara wilayah kota dengan pedesaan.

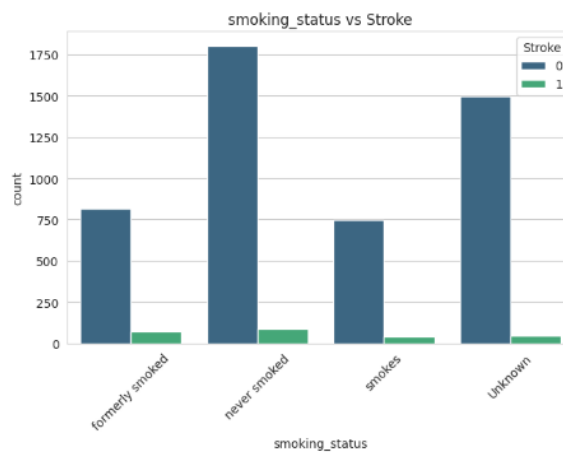


Figure 3. Hubungan Riwayat Merokok Terhadap Risiko Kejadian Stroke.

Berdasarkan faktor gaya hidup, individu dalam kategori *formerly smoked* dan *never smoked* menunjukkan frekuensi stroke yang lebih tinggi. Meskipun kelompok perokok aktif (*smokes*) juga memiliki risiko signifikan, terdapat tantangan pada kategori *Unknown* yang memiliki populasi masif. Keberadaan data *Unknown* ini tetap dipertahankan dalam analisis awal untuk menjaga integritas dataset, mengingat konversi nilai tersebut secara paksa ke nilai modus berisiko menciptakan bias interpretasi pada model prediktif.

b. Correlation Matrix (Correlation Analysis)

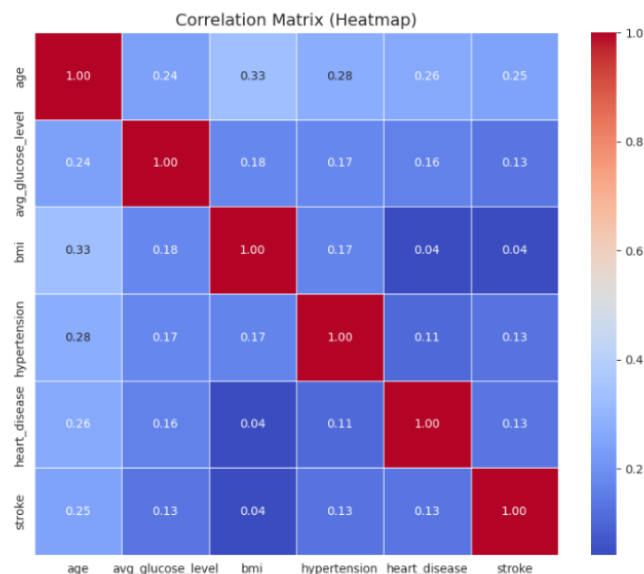


Figure 4. Correlation Matrix

variabel usia (*age*) muncul sebagai faktor yang memiliki korelasi paling konsisten dan relatif lebih kuat dibandingkan variabel lainnya terhadap risiko kesehatan. Usia memiliki korelasi tertinggi dengan BMI (0.33), diikuti oleh hipertensi (0.28), penyakit jantung (0.26), dan stroke (0.25). Hal ini mengindikasikan bahwa seiring bertambahnya usia seseorang, risiko mereka untuk memiliki BMI yang lebih tinggi, mengalami hipertensi, serta menderita penyakit jantung atau stroke cenderung meningkat secara linear dibandingkan faktor risiko lainnya.

Variabel *stroke* memiliki hubungan yang paling menonjol dengan usia (0.25), namun menunjukkan korelasi yang sangat lemah dengan BMI, yaitu hanya sebesar 0.04. Hal ini cukup menarik karena menunjukkan bahwa dalam data ini, berat badan (BMI) bukan merupakan prediktor

langsung yang kuat untuk stroke jika dibandingkan dengan faktor usia, kadar glukosa (0.13), atau keberadaan hipertensi (0.13).

Imbalanced Data Handling (*Hybrid Sampling*)

Dilakukan visualisasi menggunakan count plot untuk mengonfirmasi tingkat ketidakseimbangan rasio antara kelas 0 (non-stroke) dan kelas 1 (stroke) dalam variabel target.

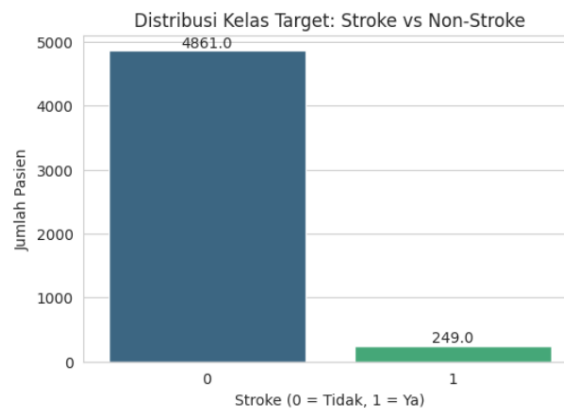


Figure 5. Distribusi Data Sebelum Hybrid Sampling

Ketidakseimbangan data (Imbalanced Data) jumlah pasien tidak stroke lebih tinggi (4861 pasien) dibandingkan dengan pasien yang mengalami stroke (249 pasien). Secara proporsional, kelas minoritas hanya mencakup sekitar 4,8% dari keseluruhan dataset, yang secara teknis dikategorikan sebagai extreme imbalance. Untuk memitigasi risiko ini dalam fase pemodelan, pendekatan yang diambil teknik Hybrid Sampling meliputi gabungan SMOTE dan Random Under Sampling

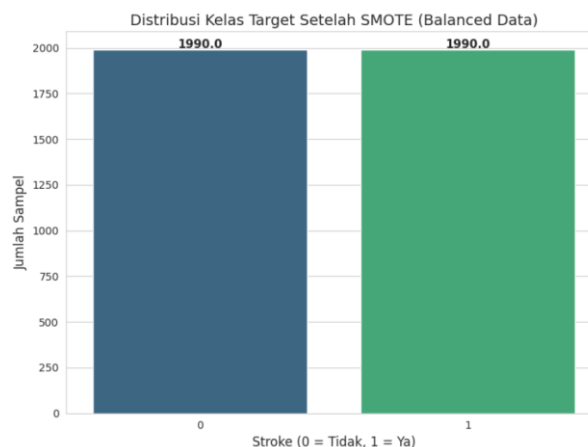


Figure 6. Distribusi Data Sesudah Hybrid Sampling
Model Performance and Evaluation

Dilakukan pemisahan antara fitur (X) dan target (Y) pada dataset, di mana target adalah kolom 'stroke' yang ingin diprediksi, sementara fitur lainnya menjadi variabel independen. Kemudian, dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80% untuk data pelatihan dan 20% untuk data pengujian menggunakan fungsi “train_test_split” dari sklearn.model_selection. Melakukan pembagian ini penting sebelum melakukan teknik SMOTE agar tidak terjadi data leakage (kebocoran data). Setelah pembagian, jumlah data pelatihan dan pengujian ditampilkan, 4088 data untuk pelatihan dan 1022 data untuk pengujian.

Model XGBoost diatur dengan beberapa hyperparameter yang telah ditentukan dalam parameter grid, seperti “learning_rate (0.05), max_depth (5), n_estimators (1000), subsample (1.0),

dan `colsample_bytree (0.8)`.” Hyperparameter ini digunakan untuk mengatur cara model XGBoost beradaptasi dengan data, di mana “`learning_rate`” mengontrol kecepatan pembelajaran, “`max_depth`” mengatur kedalaman pohon keputusan, dan “`n_estimators`” menentukan jumlah pohon dalam model. Selanjutnya, model `XGBClassifier` diinisialisasi dengan beberapa parameter tambahan, seperti “`use_label_encoder = False`” yang menonaktifkan encoder label internal, “`eval_metric = 'logloss'`” untuk menggunakan log-loss sebagai metrik evaluasi, dan “`random_state = 42`” untuk memastikan hasil yang konsisten setiap kali model dijalankan.

Table 2. Classification Report Before Hybrid Sampling

	Precision	Recall	F1 – Score	Support
0	0.95	1.00	0.97	972
1	0.00	0.00	0.00	50
Accuracy			0.95	1022
Macro Avg	0.48	0.50	0.49	1022
Weighted Avg	0.90	0.95	0.93	1022

Model klasifikasi jika tanpa menggunakan teknik Hybrid Sampling menunjukkan adanya masalah ketidakseimbangan data (imbalanced data) yang sangat ekstrem, yang berdampak buruk pada performa prediksi kelas minoritas. Berdasarkan Classification Report, model memang mencatat angka akurasi yang sekilas terlihat tinggi sebesar 95,11%, namun angka ini sangat menipu karena model hanya mampu memprediksi kelas mayoritas (0) dengan benar dan gagal total dalam mengenali kelas minoritas (1). Hal ini dibuktikan dengan nilai `precision`, `recall`, dan `f1-score` untuk kelas 1 yang seluruhnya bernilai 0.00, menandakan bahwa tidak ada satu pun sampel dari kelas minoritas yang berhasil diklasifikasikan dengan tepat.

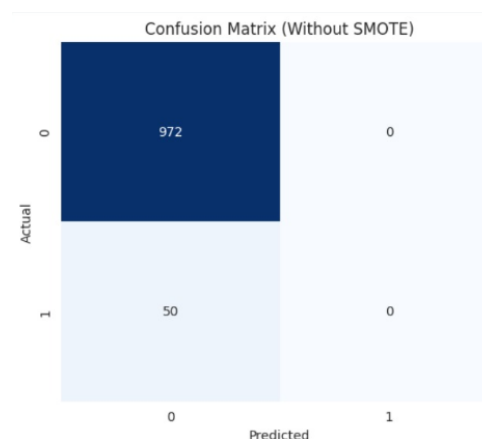


Figure 7. Confusion Matrix Before Hybrid Sampling

Visualisasi Confusion Matrix pada Figure 5 semakin mempertegas kegagalan model tersebut, di mana terlihat bahwa seluruh data uji sebanyak 1.022 sampel diprediksi masuk ke dalam kelas 0. Dari 1.022 sampel, 972 sampel terbukti benar merupakan kelas 0 (True Negative), namun terdapat 50 sampel yang seharusnya merupakan kelas 1 tetapi salah diprediksi sebagai kelas 0 (False Negative).

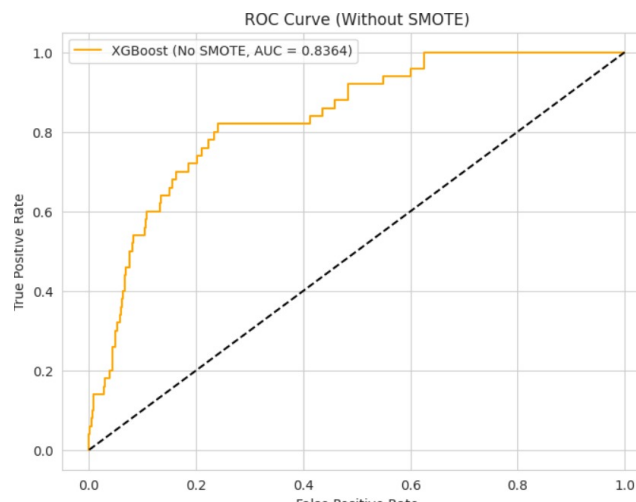


Figure 8. ROC Curve Before Hybrid Sampling

Secara matematis model memiliki kemampuan pemisahan kelas yang sangat baik pada berbagai ambang batas keputusan. Namun, nilai AUC yang tinggi ini didorong oleh dominasi kelas mayoritas (pasien non-stroke) yang mencapai 972 sampel. Meskipun area di bawah kurva luas, model ini sebenarnya mengalami bias ekstrem karena gagal memberikan prediksi yang benar pada satu pun pasien stroke di titik operasi standar.

Table 3. Classification Report After Hybrid Sampling

	Precision	Recall	F1 – Score	Support
0	0.97	0.92	0.94	972
1	0.19	0.36	0.25	50
Accuracy			0.89	1022
Macro Avg	0.58	0.64	0.60	1022
Weighted Avg	0.93	0.89	0.91	1022

Hasil evaluasi model yang menerapkan teknik Hybrid Sampling menunjukkan bahwa model memiliki akurasi sebesar 0.8943, yang berarti sekitar 89% dari prediksi model pada data uji benar. Berbeda dengan model sebelumnya yang gagal total mengenali target, model saat ini sudah mampu mengidentifikasi kelas 1 dengan nilai recall sebesar 0,36 dan f1-score sebesar 0,25. Hal ini menunjukkan bahwa Hybrid Sampling membantu model untuk keluar dari bias kelas mayoritas, sehingga model tidak lagi sekadar menebak semua data sebagai kelas 0, melainkan mulai belajar mengenali pola-pola spesifik pada pasien yang berisiko.

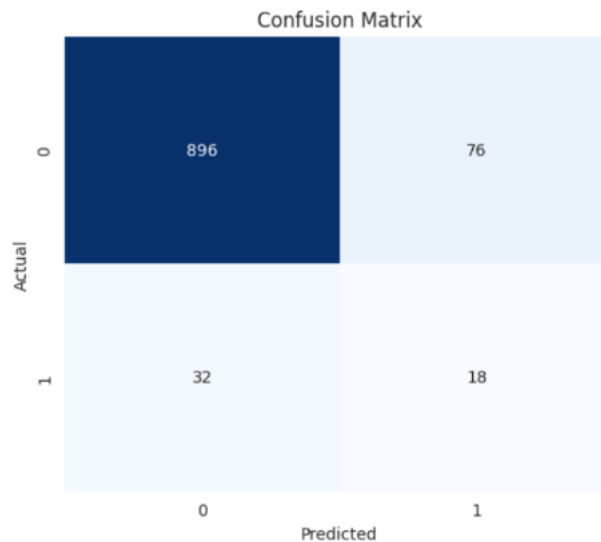


Figure 9. Confusion Matrix After Hybrid Sampling

Berdasarkan Confusion Matrix yang dihasilkan, terlihat bahwa dari 50 sampel aktual pada kelas minoritas, model berhasil memprediksi 18 sampel dengan benar sebagai True Positive. Meskipun masih terdapat 32 sampel yang salah diklasifikasikan sebagai kelas 0 (False Negative) dan adanya 76 sampel kelas mayoritas yang salah diprediksi sebagai kelas 1 (False Positive), hasil ini merupakan peningkatan kualitas deteksi yang krusial. Secara keseluruhan, meskipun precision untuk kelas minoritas masih tergolong rendah di angka 0,19, penggunaan Hybrid Sampling telah berhasil menyeimbangkan performa model sehingga lebih sensitif dan memiliki daya guna yang lebih baik dalam skenario diagnosis medis dibandingkan model tanpa penanganan ketidakseimbangan data.

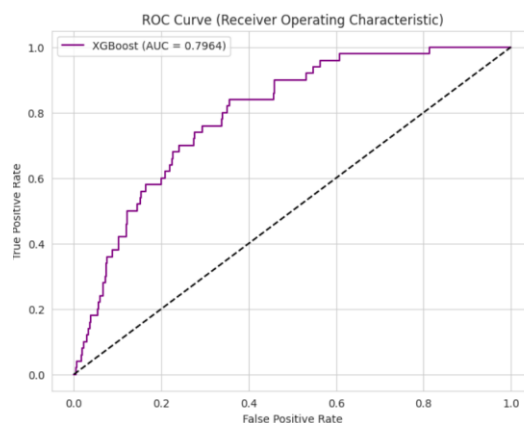


Figure 10. ROC Curve After Hybrid Sampling

Di sisi lain, ROC curve yang telah menggunakan teknik Hybrid Sampling (gabungan SMOTE dan Random UnderSampling), dengan perolehan nilai AUC sebesar 0.7964. Meskipun nilai AUC ini sedikit lebih rendah dibandingkan model sebelumnya, kurva ini mencerminkan performa model yang lebih realistis dalam menangani ketidakseimbangan data. Penurunan AUC ini merupakan upaya model untuk meningkatkan True Positive Rate, di mana model akhirnya mulai mampu mengenali 18 sampel pasien stroke yang sebelumnya tidak terdeteksi sama sekali.

Discussion

Feature Importance Analysis

Hasil analisis menunjukkan bahwa risiko stroke meningkat tajam seiring bertambahnya usia, terutama pada angka 40 hingga 80 tahun, serta sangat dipengaruhi oleh kondisi kesehatan seperti darah tinggi (hipertensi) dan kadar gula darah yang tinggi (di atas 200 mg/dL). Meskipun jumlah pasien perempuan dan pekerja swasta lebih banyak dalam data ini, risiko stroke pada laki-laki maupun penduduk desa dan kota ternyata hampir sama, sehingga lokasi tempat tinggal dan jenis kelamin bukan menjadi penyebab utama. Selain itu, berat badan yang berlebih (overweight dan obesitas) serta kebiasaan merokok terbukti memiliki hubungan erat dengan kejadian stroke. Hal ini mempertegas bahwa faktor kesehatan fisik dan gaya hidup jauh lebih berpengaruh terhadap serangan stroke dibandingkan faktor lingkungan atau identitas seseorang. Ini sejalan dengan yang diungkapkan Lannywati menyatakan bahwa faktor risiko dominan stroke dalam penelitian ini adalah umur yang semakin meningkat, jantung koroner, diabetes melitus, hipertensi, dan gagal jantung (Ghani et al., 2016).

Impact of Hybrid Sampling on Medical Data

Penerapan teknik Hybrid Sampling, yaitu Synthetic Minority Over-sampling Technique sebagai rasio terhadap Random Undersampling, memiliki dampak yang signifikan terhadap aturan masalah ketidakseimbangan kelas ekstrem, yang dialami oleh dataset ini, di mana kelas minoritas, yaitu data pasien stroke, hanya mencakup 4,8% dari total data. Teknik ini secara efektif menetralkan bias model terhadap kelas mayoritas dengan mensintesis sampel baru untuk kelas minoritas dan mengurangi jumlah elemen data kelas mayoritas, yang menciptakan distribusi yang jauh lebih seimbang. Hasil yang paling penting terjadi pada sensitivitas model, dengan nilai F1-score kelas minoritas meningkat dari 0% ke 25% setelah data dilakukan penanganan ketidakseimbangan.

Analysis of Model Performance

Berdasarkan analisis performa model, penggunaan algoritma XGBoost setelah penerapan Hybrid Sampling (gabungan SMOTE dan Random Under Sampling) menghasilkan akurasi sebesar 89,43%. Meski demikian, angka ini memang lebih rendah secara nominal daripada akurasi model sebelum Hybrid Sampling (95,11%), namun hasil ini justru lebih bagus karena model tidak lagi terperangkap oleh masalah bias terhadap kelas mayoritas sehingga mampu mengenali pola pasien yang secara spesifik resiko stroke. Berbulan model yang dibuat terhadap Confusion Matrix, model berhasil mendeteksi 18 pasien stroke sebagai true positive, yang sebelumnya sama sekali tidak terdeteksi. Didapatkan pula recall sebesar 0,36 dan AUC sebesar 0,7964. Secara umum, meskipun precision masih rendah di angka 0,19, model ini dianggap lebih realistis dan memiliki kegunaan medis yang lebih baik sehingga risiko pasien stroke yang terdeteksi menjadi lebih minim.

KE SIMPULAN

Analisis EDA mengungkapkan bahwa faktor seperti usia, kadar glukosa, dan BMI berkorelasi nyata dengan penyakit stroke. Penerapan Hybrid Sampling (gabungan SMOTE dan Random Under Sampling) berhasil menyeimbangkan distribusi kelas sehingga memudahkan model mempelajari fitur-fitur kelas minoritas tanpa didominasi oleh kelas mayoritas. dibuktikan dari meningkatnya angka f1-score kelas 1 yang awalnya 0 menjadi 0.25 atau 25 %. Berdasarkan hasil implementasi dan evaluasi algoritma XGBoost untuk klasifikasi risiko penyakit stroke, dapat ditarik kesimpulan bahwa pada dataset ini memiliki ketidakseimbangan kelas yang sangat besar, dengan kelas minoritas (strok) hanya 4,8% (249 pasien) dan kelas mayoritas (non-strok) 95,2% (4861 persen). Model menunjukkan akurasi yang cukup baik secara keseluruhan (89,43%), namun masih menghadapi tantangan dalam mengidentifikasi kasus stroke secara akurat, dengan recall sebesar 0,36 dan precision 0,19 untuk kelas minoritas. Hal ini menunjukkan bahwa meskipun akurasi tinggi model cenderung bias terhadap kelas mayoritas dan rendah dalam mendeteksi pasien berisiko stroke.

DAFTAR PUSTAKA

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>.
- Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors*, 22(13), 4670. <https://doi.org/10.3390/s22134670>
- Faisal, A., & Subekti, A. (2022). Deep neural network untuk prediksi stroke menggunakan SMOTE Tomek. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 8(1).
- Ghani, L., Miwardja, L. K., & Delima. (2016). Faktor risiko dominan penderita stroke di Indonesia. *Buletin Penelitian Kesehatan*, 44(1), 49-58. <https://www.google.com/search?q=https://doi.org/10.22435/bpk.v44i1.5625>.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.
- Kementerian Kesehatan Republik Indonesia. (2019). *Laporan nasional Riskesdas 2018*. Jakarta: Badan Penelitian dan Pengembangan Kesehatan.
- Kianie, K., & Cenggonono, M. (2025). Pengaruh faktor risiko dan gaya hidup terhadap risiko stroke. *Jurnal Pendidikan Tambusai*, 9(1).
- Nabila, H. A., & Pamungkas, E. W. (2023). Perbandingan algoritma machine learning: SVM, Random Forest, dan XGBoost untuk prediksi stroke. *RABIT: Jurnal Teknologi dan Sistem Informasi*, 8(2).
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
- World Health Organization. (2023). *Stroke (cerebrovascular accident)*. Jenewa: World Health Organization.