JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 274~281
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608

Komparasi Algoritma LightGBM, SVM, dan Logistic Regression dalam Memprediksi Penyakit Stroke

[1]Bryant Steven Aritonang, [2]Umniy Salamah
[1,2] Fakultas Ilmu Komputer, Universitas Mercu Buana, Indonesia
[1]41521010119@student.mercubuana.ac.id; [2]umniy.salamah@mercubuana.ac.id;

Article Info

Article history:

Received, 2025-01-04
Revised, 2025-01-14
Accepted, 2025-01-31

Kata Kunci:

Stroke
LightGBM
SVM
Logistik Regresi

ABSTRAK

Stroke merupakan penyakit serius yang dapat menyebabkan kecacatan atau kematian akibat gangguan aliran darah ke otak. Penelitian ini bertujuan untuk membandingkan tiga algoritma machine learning, yaitu LightGBM, Support Vector Machine (SVM), dan Logistic Regression dalam memprediksi risiko stroke. Dataset yang digunakan berjumlah 5110 baris dengan 12 atribut, yang mencakup informasi demografi dan riwayat kesehatan pasien. Proses penelitian dimulai dengan preprocessing data, diikuti dengan pembagian data menjadi data latih dan data uji. Selanjutnya, model dilatih menggunakan ketiga algoritma dan dievaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil analisis menunjukkan bahwa Logistic Regression memiliki performa terbaik secara keseluruhan dengan keseimbangan antara deteksi kasus stroke dan mengidentifikasi individu yang sehat. SVM menunjukkan hasil yang stabil dengan keseimbangan antara recall dan precision, sementara LightGBM, meskipun memiliki akurasi tinggi, kurang efektif dalam mendeteksi kasus stroke. Penelitian ini menyimpulkan bahwa Logistic Regression adalah model yang paling sesuai untuk memprediksi risiko stroke, meskipun SVM dapat menjadi alternatif yang baik.

Keywords:

Stroke
LightGBM
SVM
Logistic Regression

ABSTRACT

Stroke is a serious condition that can lead to disability or death due to disrupted blood flow to the brain. This study aims to compare three machine learning algorithms: LightGBM, Support Vector Machine (SVM), and Logistic Regression, in predicting the risk of stroke. The dataset used contains 5110 rows with 12 attributes, including demographic information and health history. The research process began with data preprocessing, followed by splitting the data into training and testing sets. Models were then trained using the three algorithms and evaluated using accuracy, precision, recall, and F1-score metrics. The analysis results indicate that Logistic Regression performed the best overall, providing a balance between detecting stroke cases and identifying healthy individuals. SVM showed stable results with a balance between recall and precision, while LightGBM, despite high accuracy, was less effective in detecting stroke cases. The study concludes that Logistic Regression is the most suitable model for predicting stroke risk, though SVM can be a good alternative.

This is an open access article under the CC BY-SAlicense.

Penulis Korespondensi:

Bryant Steven Aritonang,
Program Studi Teknik Informatika,
Universitas Mercu Buana,
Email: 41521010119@student.mercubuana.ac.id

1. PENDAHULUAN

Stroke adalah kondisi neurologis serius yang terjadi ketika aliran darah ke bagian otak terganggu[1], [2], [3], [4], baik karena perdarahan atau sumbatan[5]. Ini dapat mengakibatkan kerusakan otak yang menyebabkan gejala seperti kelemahan otot, kesulitan berbicara, gangguan penglihatan, dan bahkan kehilangan kesadaran yang dapat menyebabkan cacat atau kematian[6]. Berdasarkan sejarahnya, Hippocrates pertama kali mendokumentasikan tentang stroke sekitar 400 tahun sebelum masehi, menggunakan istilah "Apoplexy" untuk menggambarkan kondisi ini pada pasien. Pada abad ke-17, Johann Jakob Wepfer dan Thomas Willis

melanjutkan penelitian ini, diikuti oleh peneliti lain seperti Giovanni Battista Morgagni, Charles Miller Fisher, dan Rudolf Virchow. Hingga kini, penelitian dan pengembangan terkait stroke terus berlanjut[7].

Stroke menjadi salah satu penyebab utama kematian di seluruh dunia[8], baik di negara maju maupun berkembang[9], [10]. Data dari World Stroke Organization menunjukkan bahwa setiap tahunnya terjadi sekitar 13,7 juta kasus baru stroke, yang menyebabkan sekitar 5,5 juta kematian [11].

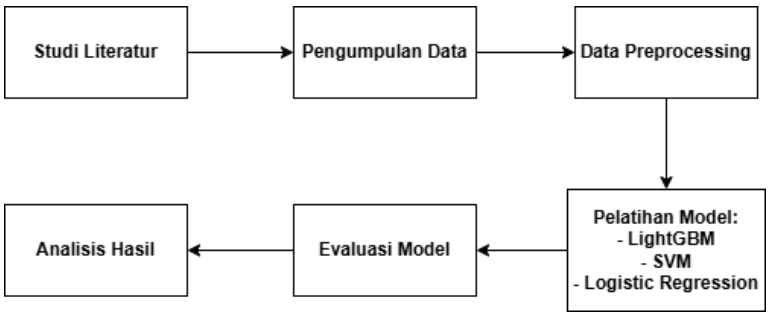
Faktor resiko penyebab seseorang terkena penyakit stroke dibagi menjadi dua yaitu faktor yang tidak dapat dimodifikasi (usia, jenis kelamin, ras, genetik atau riwayat keluarga) dan juga faktor yang dapat dimodifikasi (hipertensi atau tekanan darah, kelebihan berat badan, gula darah atau diabetes, kolesterol, merokok, alkohol, polusi, dll)[12]. Meskipun ada upaya pencegahan dan penanganan dini, stroke tetap menjadi salah satu penyebab utama kematian dan cacat di seluruh dunia[13]. Salah satu permasalahan utama dalam penanganan stroke adalah keterlambatan untuk melakukan diagnosis. Untuk mengatasi masalah tersebut, penelitian ini bertujuan untuk melakukan komparasi beberapa algoritma machine learning yang berbeda dalam melakukan prediksi penyakit stroke.[14]

Terdapat beberapa penelitan terdahulu terkait penyakit stroke menggunakan berbagai data dan metode. Sebagai contoh, Felix Indra Kurniadi dan Pramitha Dwi Larasati melakukan penelitian dengan judul "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke", di mana algoritma LightGBM mencapai akurasi 98%[15]. Kemudian, Ulfa Amelia, Jamaludin Indra, dan Anis Fitri Nur Masruriyah melakukan penelitian dengan judul "IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK PREDIKSI PENYAKIT STROKE DENGAN ATRIBUT BERPENGARUH", di mana algoritma SVM mencapai akurasi 100% [16]. Peneliti lainnya, seperti Yufis Azhar, Aidia Khoiriyah Firdausy, dan Putri Juli Amelia, membahas Logistik Regresi dalam penelitian "Perbandingan Algoritma Klasifikasi Data Mining untuk Prediksi Penyakit Stroke", dengan hasil akurasi sebesar 73,48% [17].

Berdasarkan penjelasan latar belakang di atas, penelitian selanjutnya akan membandingkan metode LightGBM, Support Vector Machine (SVM), dan Logistic Regression[18] dalam memprediksi penyakit stroke. Data yang digunakan berasal dari sumber www.kaggle.com. Diharapkan melalui perbandingan algoritma tersebut dapat memberikan prediksi yang lebih akurat dan efisien terhadap kemungkinan terjadinya stroke pada individu. Dengan demikian, peneliti dapat mengevaluasi dan membandingkan kinerja ketiga algoritma ini dalam memprediksi risiko stroke dengan menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan f1-score.

2. METODE PENELITIAN

Dalam penelitian ini, alur yang digunakan akan mengikuti langkah-langkah yang telah diterapkan oleh peneliti, yang dapat dilihat pada bagan di bawah ini:



Gambar 1 Tahapan Penelitian

Metode penelitian yang dilakukan pada penelitian ini mengikuti alur sistematis yang melibatkan beberapa tahap, dimulai dari pengumpulan data hingga analisis hasil. Data yang digunakan dalam penelitian ini diambil dari platform Kaggle. Dataset tersebut memiliki 5110 baris dan terdiri dari 12 atribut, yaitu: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, dan stroke. Atribut-atribut tersebut mencakup informasi demografi, riwayat kesehatan, serta status terkait risiko stroke pada individu.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 274~281
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Gambar 2 Dataframe

Tahap berikutnya adalah data preprocessing. Pada tahap ini, berbagai teknik diterapkan untuk membersihkan dan mempersiapkan data agar sesuai dengan kebutuhan analisis. Proses ini meliputi penanganan data yang hilang, normalisasi atribut numerik, pengkodean data kategori menjadi format numerik agar dapat diterima oleh model. Setelah preprocessing selesai, dataset dibagi menjadi dua subset: **80% untuk data latih** (training data) dan **20% untuk data uji** (testing data). Pembagian ini dilakukan untuk memastikan model dapat dilatih dengan baik dan dievaluasi secara obyektif.

Tahapan selanjutnya adalah pelatihan model, dimana tiga algoritma berbeda digunakan, yaitu LightGBM, Support Vector Machine (SVM), dan Logistic Regression. LightGBM digunakan karena kemampuannya yang efisien dalam menangani dataset besar dengan berbagai tipe fitur. SVM dipilih untuk kemampuan klasifikasinya yang kuat pada data berdimensi tinggi, sedangkan Logistic Regression digunakan sebagai baseline model karena kesederhanaannya dan interpretabilitas yang baik.

Tahap evaluasi model dilakukan untuk mengukur performa masing-masing model dalam memprediksi risiko stroke. Metode evaluasi mencakup metrik seperti akurasi, precision, recall, dan F1-score. Perbandingan hasil evaluasi ini membantu dalam menentukan model yang paling optimal untuk tugas klasifikasi stroke.

Tahap terakhir adalah analisis hasil, di mana hasil evaluasi dari model-model tersebut dianalisis lebih lanjut untuk mendapatkan wawasan mengenai pola-pola penting dalam data. Hasil analisis ini memberikan informasi yang berguna dalam memahami faktor-faktor yang berkontribusi terhadap risiko stroke serta memvalidasi efektivitas pendekatan penelitian yang digunakan.

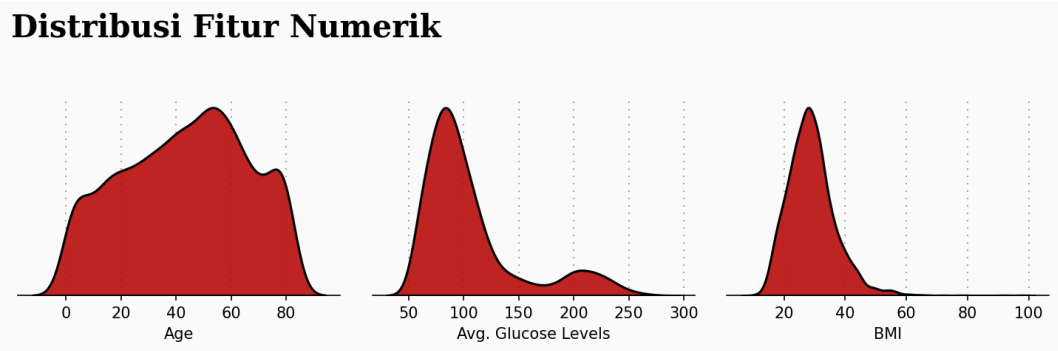
3. HASIL DAN ANALISIS

Sebagai hasil dari penelitian ini, dilakukan komparasi algoritma untuk memprediksi kemungkinan seseorang mengalami penyakit stroke. Metode yang digunakan melibatkan algoritma LightGBM, SVM, dan Logistic Regression dalam menganalisis data pasien berdasarkan fitur-fitur yang relevan.

Gambar 3 Distribusi Penyakit Stroke

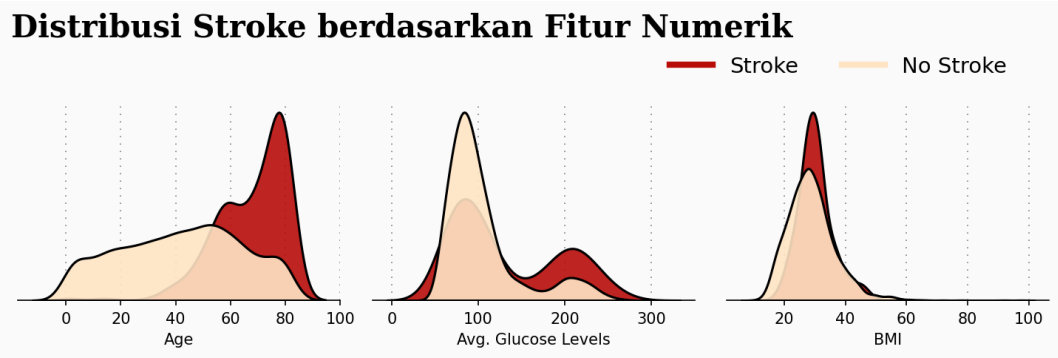
Berdasarkan distribusi data, hanya sekitar 5% dari total sampel yang termasuk dalam kelas minoritas (mengalami stroke), sementara 95% sisanya berada dalam kelas mayoritas (tidak mengalami stroke). Ketidakseimbangan ini menunjukkan adanya *imbalance data*, di mana dominasi kelas mayoritas dapat menyebabkan model prediksi cenderung mengabaikan kelas minoritas. Akibatnya, model lebih fokus pada prediksi kelas mayoritas untuk mencapai akurasi keseluruhan yang tinggi, namun berpotensi mengurangi performa dalam mendeteksi kasus stroke.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 274~281
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608



Gambar 4 Distribusi fitur Numerik

Analisis distribusi fitur numerik, seperti *age, bmi, dan avg_glucose_leve*l, dilakukan untuk menggali pola dan karakteristik data yang berpotensi memengaruhi kinerja model dalam proses prediksi. Dari hasil analisis, terlihat bahwa distribusi usia cenderung mengikuti pola normal. Namun, pada variabel *avg_glucose_level* dan *bmi* ditemukan adanya positive skewness, yang mengindikasikan keberadaan nilai ekstrem (*outlier*) pada kedua variabel tersebut.



Gambar 5 Distribusi Stroke berdasarkan Fitur Numerik

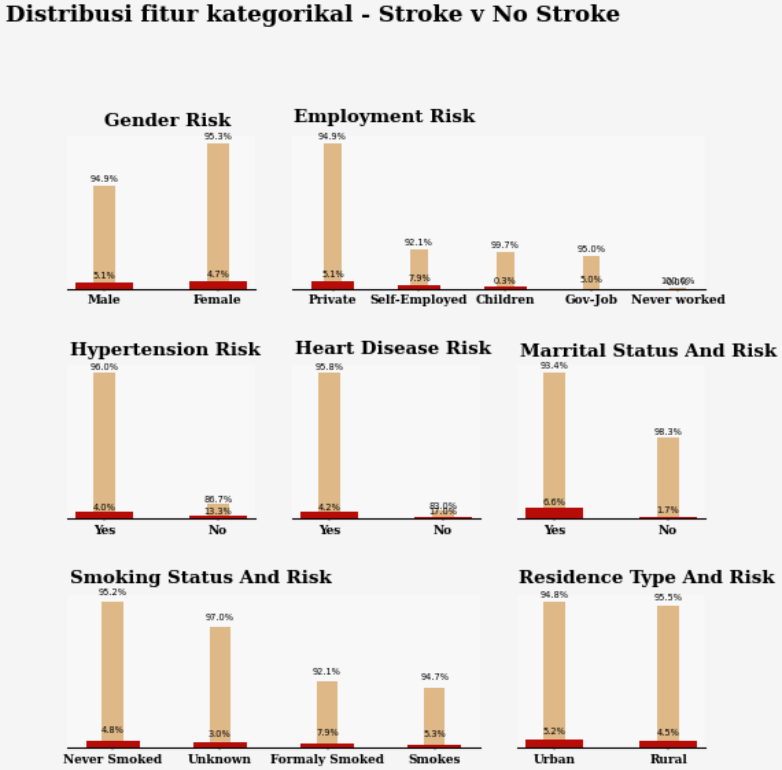
Usia merupakan faktor utama yang memengaruhi risiko stroke, dengan prevalensi yang meningkat pada kelompok usia lanjut. Kadar glukosa darah tinggi juga memiliki hubungan signifikan dengan kejadian stroke, sedangkan BMI memberikan pengaruh moderat, terutama dalam konteks obesitas. Penelitian ini kemudian difokuskan pada analisis lebih mendalam terhadap atribut usia untuk memahami hubungan antara kelompok umur tertentu dengan prevalensi stroke.

Gambar 6 Hubungan antara Usia dan Penyakit Stroke

Risiko stroke meningkat seiring bertambahnya usia, dengan percepatan signifikan mulai usia 60 tahun ke atas dan mencapai puncaknya setelah 80 tahun. Pada usia muda, terutama di bawah 40 tahun, risiko tetap rendah dan stabil. Hal ini menegaskan bahwa usia adalah faktor utama dalam risiko stroke, menekankan pentingnya pencegahan sejak dini untuk mengurangi risiko di masa depan.

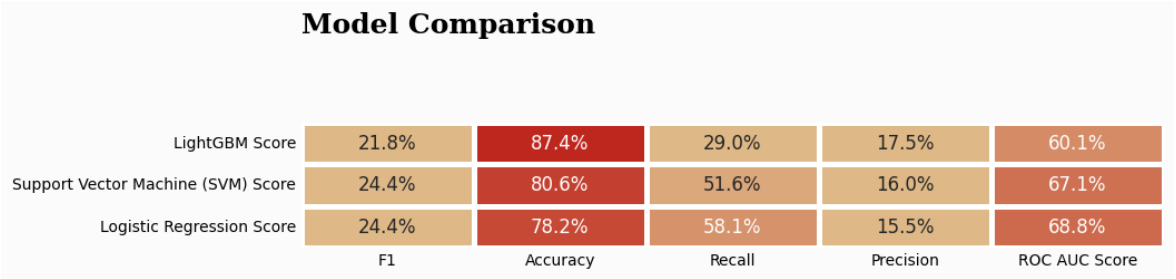
JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 274~281
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608

Selanjutnya dilakukan analisis terkait fitur kategorikal yang digunakan untuk melihat berbagai faktor demografis, sosial, dan kesehatan terhadap resiko stroke. Proses ini membantu menentukan variabel yang paling signifikan dan mendukung pengembangan model prediktif yang lebih akurat serta langkah pencegahan yang efektif.



Gambar 7 Distribusi Fitur Kategorikal terhadap Penyakit Stroke

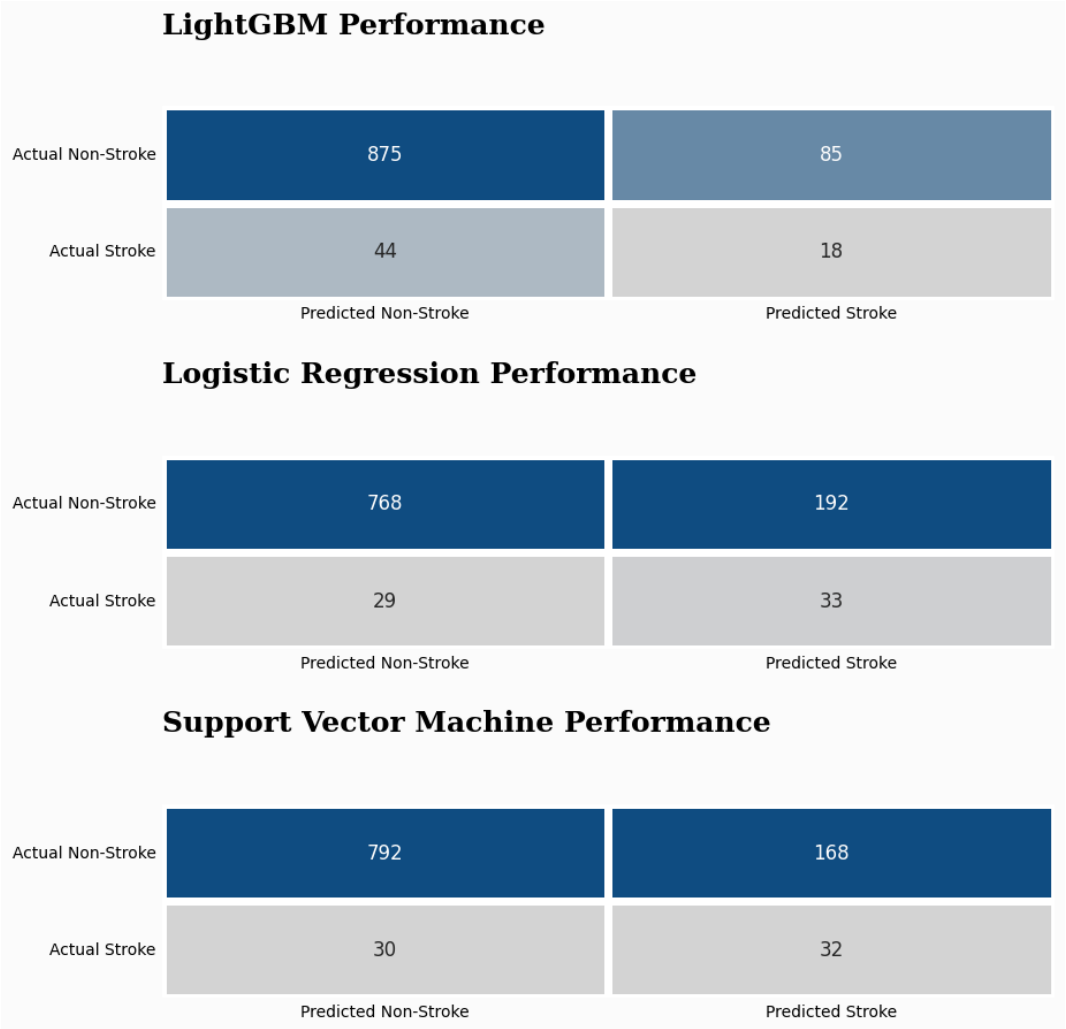
Proporsi yang lebih tinggi untuk kategori "Yes" pada faktor-faktor tertentu menunjukkan bahwa hipertensi, penyakit jantung, jenis pekerjaan (Self-Employed), dan status merokok signifikan berkontribusi pada peningkatan risiko stroke. Sebaliknya, gender, status menikah, dan lokasi tempat tinggal memiliki dampak yang relatif kecil.



Gambar 8 Hasil Komparasi Model

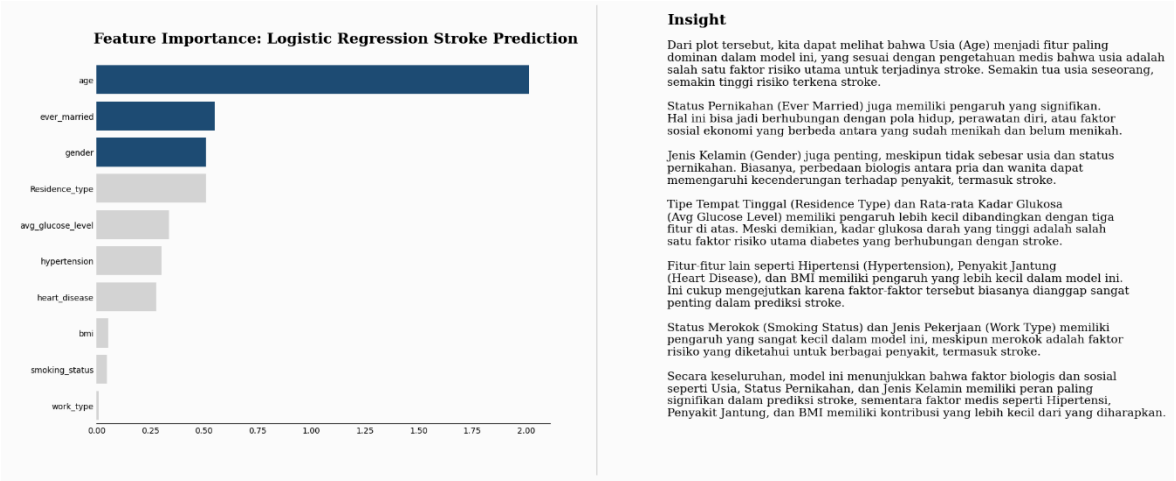
Dari hasil komparasi algoritma, Logistic Regression menunjukkan performa terbaik secara keseluruhan karena mampu memberikan keseimbangan antara kemampuan mendeteksi kelas positif dan meminimalkan kesalahan. SVM menjadi alternatif yang cukup baik dengan kekuatan pada pengurangan kesalahan prediksi positif palsu, meskipun sensitivitasnya terhadap kelas minoritas sedikit lebih rendah. Sementara itu, LightGBM meskipun memiliki akurasi tinggi, kurang cocok untuk skenario ini karena cenderung hanya memprediksi kelas mayoritas, sehingga mengabaikan kelas minoritas. Berdasarkan hasil ini, Logistic Regression direkomendasikan sebagai model yang paling sesuai untuk digunakan.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 274~281
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608



Gambar 9 Confusion Matrix

Meskipun LightGBM unggul dalam mengidentifikasi individu tanpa stroke, tetapi kurang efektif dalam mendeteksi kasus stroke yang sebenarnya. Logistic Regression lebih baik dalam mendeteksi kasus stroke, namun cenderung menghasilkan lebih banyak kesalahan prediksi pada individu tanpa stroke. SVM menawarkan keseimbangan yang lebih baik antara kesalahan prediksi dan kemampuan deteksi, meskipun performanya masih di bawah Logistic Regression dalam mengidentifikasi kasus stroke.



Gambar 10 Features Importance dan Insight yang diperoleh

4. KESIMPULAN

Berdasarkan hasil eksperimen prediksi penyakit stroke dengan menggunakan dataset dari Kaggle yang berjudul 'Stroke Prediction Dataset' oleh Fedesoriano, model Logistic Regression menunjukkan performa terbaik dengan keseimbangan yang baik antara deteksi kasus stroke dan membedakan kelas negatif, menghasilkan hasil yang lebih optimal. SVM memberikan hasil yang stabil dengan keseimbangan antara Recall dan Precision, meskipun sedikit kurang sensitif dalam mendeteksi kasus stroke dibandingkan Logistic Regression. Sementara itu, LightGBM, meskipun memiliki akurasi tinggi, menunjukkan kelemahan dalam mendeteksi kelas positif, menjadikannya kurang efektif untuk kasus ini.

UCAPAN TERIMA KASIH

Terima kasih saya sampaikan kepada Fakultas Ilmu Komputer Universitas Mercu Buana terkhususnya pada Program Studi Teknik Informatika yang telah membantu dalam penulisan artikel ini.

REFERENSI

[1] D. Ismafillah, T. Rohana, and Y. Cahyana, "Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke," *Jurnal Riset Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.

[2] G. C. Okoye and E. U. Umeh, "Predicting Functional Outcome After Ischemic Stroke Using Logistic Regression and Machine Learning Models," *Earthline Journal of Mathematical Sciences*, pp. 133–150, Nov. 2023, doi: 10.34198/ejms.14124.133150.

[3] G. Sailasya and G. L. Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms." [Online]. Available: www.ijacsa.thesai.org

[4] C. Author *et al.*, "Brain Tumor Detection and Classification Using Fine-Tuned CNN with ResNet50 and EfficientNet," *International Journal of Informatics and Computation (IJICOM)*, vol. 6, no. 1, 2024, doi: 10.35842/ijicom.

[5] M. N. Maskuri, K. Sukerti, and R. M. Herdian 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.

[6] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," *Revue d'Intelligence Artificielle*, vol. 34, no. 6, pp. 753–761, Dec. 2020, doi: 10.18280/RIA.340609.

[7] Y. Xue *et al.*, "The Prediction Models for High-Risk Population of Stroke Based on Logistic Regressive Analysis and Lightgbm Algorithm Separately," 2022. [Online]. Available: https://creativecommons.org/licenses/by-nc/4.0/

[8] K. Pallavi and Saravananthirunavakarasu, "Classification Of Stroke Disease Using Machine Learning Algorithms," 2022. [Online]. Available: www.ijcrt.org

[9] A. Setiawan, R. Febrio Waleska, M. Adji Purnama, and L. Efrizoni, "KOMPARASI ALGORITMA K-NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), DAN DECISION TREE DALAM KLASIFIKASI PENYAKIT STROKE," 2024. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

[10] K. Seo *et al.*, "Forecasting the walking assistance rehabilitation level of stroke patients using artificial intelligence," *Diagnostics*, vol. 11, no. 6, Jun. 2021, doi: 10.3390/diagnostics11061096.

[11] S. Wulandari, Y. Isro, T. Susanti Teknik Informatika, I. Teknologi Pagar Alam Jalan Masik Siagim No, K. Dalo, and K. Dempo Tengah, "OPTIMALISASI PREDIKSI PENYAKIT STROKE MENGGUNAKAN ALGORITMA DEEP LEARNING," 2024.

[12] J. Yu, S. Park, S. H. Kwon, C. M. B. Ho, C. S. Pyo, and H. Lee, "AI-based stroke disease prediction system using real-time electromyography signals," *Applied Sciences (Switzerland)*, vol. 10, no. 19, Oct. 2020, doi: 10.3390/app10196791.

[13] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," *Jurnal Infomedia: Teknik Informatika*.

[14] T. Hamaguchi *et al.*, "Support Vector Machine-Based Classifier for the Assessment of Finger Movement of Stroke Patients Undergoing Rehabilitation," *J Med Biol Eng*, vol. 40, no. 1, pp. 91–100, Feb. 2020, doi: 10.1007/s40846-019-00491-w.

[15] F. I. Kurniadi and P. D. Larasati, "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke."

[16] Ulfa Amelia, Jamaludin Indra, and Anis Fitri Nur Masruriyah, "IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK PREDIKSI PENYAKIT STROKE DENGAN ATRIBUT BERPENGARUH," *Scientific Student Journal for Information, Technology and Science*, vol. III, pp. 2715–2766, Jul. 2022.

[17] Y. Azhar, A. Khoiriyah Firdausy, and P. J. Amelia, "SINTECH Journal | 191 Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke", [Online]. Available: https://doi.org/10.31598

[18] W. Sardjono, A. Retnowardhani, R. Emil Kaburuan, and A. Rahmasari, "Artificial intelligence and big data analysis implementation in electronic medical records," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2021, pp. 231–237. doi: 10.1145/3512576.3512618.