JT : Jurnal Teknik
P-ISSN: 2302-8734
E-ISSN: 2581-0006

Vol. 12 No. 02 Th. 2023
Halaman : 72 - 80.
07 Oktober 2023

KLASIFIKASI DAN EVALUASI PERFORMA MODEL RANDOM FOREST UNTUK PREDIKSI STROKE

Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction

[1]Muhamad Fadli, [2]Rizal Adi Saputra
[1]Universitas Halu Oleo
[2]Universitas Halu Oleo
e-mail: fadlimuham809@gmail.com

Received: 13 Juli 2023 *Accepted: 07 Oktober 2023*

Abstract

This research aims to develop and evaluate the performance of a classification model using the Random Forest algorithm for stroke prediction. The study utilizes clinical data and risk factors collected from patients involved in previous stroke research. The data includes information such as age, gender, family history, blood pressure, cholesterol, smoking habits, and more. The research employs an experimental method, where data is collected, processed, and divided into training and testing sets. The Random Forest model is trained using the training data and used to make predictions on the testing data. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1 score, while a confusion matrix is used to visualize the prediction results. The results and discussion of this research demonstrate that the Random Forest model performs well in predicting stroke, achieving an accuracy of 93.72%, precision of 91.32%, recall of 96.59%, and an F1 score of 93.88%. These findings indicate that machine learning techniques, such as Random Forest, can be effectively used as a method for stroke prediction based on clinical data and risk factors. This research provides new insights into the development of more accurate prediction methods to support stroke diagnosis and prevention using machine learning approaches.

***Keywords**: Stroke, Machine Learning, Random Forest*

Abstrak

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi performa model klasifikasi menggunakan algoritma Random Forest untuk prediksi stroke. Studi ini menggunakan data klinis dan faktor risiko yang dikumpulkan dari pasien yang terlibat dalam penelitian stroke sebelumnya. Data ini mencakup informasi seperti usia, jenis kelamin, riwayat keluarga, tekanan darah, kolesterol, kebiasaan merokok, dan banyak lagi. Metode eksperimen digunakan dalam penelitian ini, di mana data dikumpulkan, diproses, dan dibagi menjadi data pelatihan dan pengujian. Model Random Forest dilatih dengan data pelatihan dan digunakan untuk melakukan prediksi pada data pengujian. Evaluasi performa model dilakukan menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan skor F1, serta matriks konfusi digunakan untuk memvisualisasikan hasil prediksi. Hasil dan pembahasan penelitian ini menunjukkan bahwa model Random Forest memiliki performa

yang baik dalam memprediksi stroke dengan akurasi sebesar 93,6%, presisi sebesar 91,4%, *recall* sebesar 96,1%, dan F1-Score sebesar 93,7%. Hasil ini menunjukkan bahwa teknik *machine learning* seperti Random Forest dapat digunakan sebagai metode yang efektif dalam prediksi stroke berdasarkan data klinis dan faktor risiko. Penelitian ini memberikan wawasan baru dalam pengembangan metode prediksi yang lebih akurat untuk mendukung diagnosis dan pencegahan stroke menggunakan pendekatan machine learning.

Kata Kunci: *Stroke, Machine Learning, Random Forest*

PENDAHULUAN

Stroke adalah penyakit pada otak berupa gangguan fungsi saraf lokal dan atau global, yang muncul mendadak, progresif, dan cepat. Gangguan fungsi saraf pada stroke disebabkan oleh gangguan peredaran darah otak non traumatik.

Stroke merupakan salah satu penyakit yang mempengaruhi jutaan orang di seluruh dunia dan menjadi penyebab utama kecacatan dan kematian. Pengidentifikasian dini faktor risiko dan prediksi potensi stroke menjadi sangat penting dalam upaya pencegahan, pengobatan, dan pengelolaan penyakit ini. Dalam beberapa tahun terakhir, teknik-teknik machine learning telah memperoleh perhatian besar dalam bidang medis, khususnya dalam prediksi dan diagnosis penyakit berdasarkan data klinis.

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi performa model klasifikasi menggunakan algoritma Random Forest untuk prediksi stroke. Random Forest adalah algoritma machine learning yang telah terbukti efektif dalam menangani masalah klasifikasi pada dataset kompleks. Model Random Forest menggabungkan beberapa pohon keputusan secara ensemble, di mana setiap pohon memberikan suara untuk keputusan akhir.

Dataset yang digunakan dalam penelitian ini berisi data klinis dan faktor risiko yang dikumpulkan dari pasien yang terlibat dalam penelitian stroke sebelumnya. Fitur-fitur ini mencakup informasi seperti usia, jenis kelamin, riwayat keluarga, tekanan darah, kolesterol, kebiasaan merokok, dan banyak lagi. Dengan menggunakan dataset ini, kami ingin melatih model Random Forest untuk mengklasifikasikan pasien sebagai "stroke" atau "non-stroke" berdasarkan fitur-fitur ini.

Pendekatan yang kami terapkan meliputi praproses data, termasuk pengisian nilai yang hilang, transformasi variabel kategorikal menjadi numerik menggunakan teknik LabelEncoder, dan pembagian data menjadi data pelatihan dan pengujian. Data pelatihan digunakan untuk melatih model Random Forest, sedangkan data pengujian digunakan untuk mengevaluasi performa model yang dihasilkan.

Evaluasi performa model dilakukan menggunakan berbagai metrik evaluasi yang umum digunakan dalam klasifikasi, seperti akurasi, presisi, recall, dan skor F1. Akurasi mengukur sejauh mana model dapat mengklasifikasikan dengan benar, sedangkan presisi mengukur sejauh mana model memberikan prediksi yang benar untuk kelas positif. Recall mengukur sejauh mana model dapat mendeteksi dengan benar kelas positif, sedangkan skor F1 adalah penggabungan antara presisi dan recall.

Selain itu, kami juga menggunakan matriks konfusi untuk memvisualisasikan hasil prediksi dan klasifikasi yang dilakukan oleh model. Matriks konfusi memberikan gambaran tentang seberapa baik model dapat mengklasifikasikan sampel ke dalam kelas positif dan negatif, serta seberapa sering model mengalami kesalahan dalam prediksi.

Penelitian ini diharapkan dapat memberikan wawasan baru tentang kemampuan model Random Forest dalam prediksi stroke, serta memberikan kontribusi pada bidang kesehatan dalam pengembangan metode prediksi yang lebih akurat. Dengan menggunakan teknik machine learning, diharapkan dapat meningkatkan kemampuan prediksi dini stroke dan memberikan arahan yang lebih baik dalam pengambilan keputusan klinis.

Melalui jurnal ini, kami berharap dapat memberikan pemahaman yang lebih baik tentang penggunaan model Random Forest dalam prediksi stroke, serta menginspirasi penelitian dan pengembangan lebih lanjut dalam bidang machine learning dan kesehatan. Studi ini juga dapat menjadi landasan untuk penelitian lebih lanjut dalam menggali potensi machine learning dalam mendukung diagnosis dan prediksi penyakit lainnya.

METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen untuk menguji performa model Random Forest dalam prediksi stroke. Data dikumpulkan, diproses, dan dibagi menjadi data pelatihan dan pengujian. Model Random Forest dilatih dengan data pelatihan dan digunakan untuk melakukan prediksi pada data pengujian. Performa model dievaluasi menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-*score*, serta hasilnya disajikan dalam bentuk *confusion matrix*.

Pengumpulan Data:

Data klinis dan faktor risiko yang berkaitan dengan stroke dikumpulkan dari penelitian sebelumnya. Data ini mencakup informasi seperti usia, jenis kelamin, riwayat keluarga, tekanan darah, kolesterol, kebiasaan merokok, dan variabel lain yang relevan dalam prediksi stroke.

Praproses Data:

Data yang dikumpulkan melalui langkah pertama dianalisis dan dipersiapkan untuk pemodelan. Langkah praproses data mencakup penghapusan nilai yang hilang, pengkodean variabel kategorikal menjadi numerik menggunakan teknik seperti LabelEncoder, normalisasi variabel numerik, dan pengelompokan data jika diperlukan.

Pembagian Data:

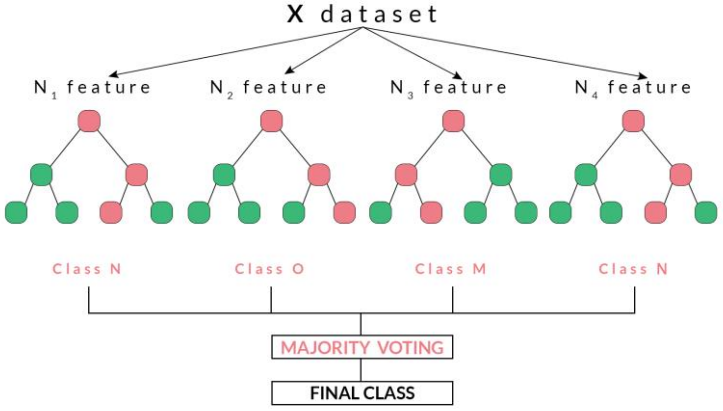
Setelah praproses data, dataset dibagi menjadi data pelatihan dan data pengujian. Umumnya, sebagian besar data digunakan untuk pelatihan model, sementara sisanya digunakan untuk pengujian dan evaluasi.

Oversampling dengan SMOTE:

Karena dataset stroke cenderung tidak seimbang, di mana jumlah sampel stroke mungkin lebih sedikit daripada jumlah sampel non-stroke, teknik oversampling menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*) dapat diterapkan untuk mengatasi masalah ketidakseimbangan ini. Metode SMOTE menciptakan sampel sintetis untuk kelas minoritas sehingga jumlah sampel stroke dapat diimbangi dengan jumlah sampel non-stroke.

Pelatihan Model:

Model Random Forest dilatih menggunakan data pelatihan yang telah disiapkan. Algoritma Random Forest digunakan karena kemampuannya dalam menangani dataset yang kompleks dan memberikan hasil yang baik dalam klasifikasi. Pohon keputusan secara ensemble digunakan dalam model ini, di mana setiap pohon memberikan suara untuk keputusan akhir.



Gambar 1 Algoritma Random Forest

Evaluasi Model:

Setelah pelatihan, model yang dihasilkan dievaluasi menggunakan data pengujian yang terpisah. Metrik evaluasi seperti akurasi, presisi, recall, dan skor F1 digunakan untuk mengukur performa model dalam mengklasifikasikan stroke dan *non*-stroke. Selain itu, matriks konfusi juga digunakan untuk memvisualisasikan hasil prediksi dan klasifikasi yang dilakukan oleh model.

Analisis Hasil:

Hasil evaluasi model dianalisis untuk mengevaluasi performa model Random Forest dalam prediksi stroke. Kesimpulan dan temuan penting diambil berdasarkan metrik evaluasi yang digunakan dan interpretasi matriks konfusi. Hasil ini dapat digunakan sebagai dasar untuk menyimpulkan keefektifan model dalam klasifikasi stroke.

Diskusi dan Kesimpulan:

Hasil penelitian dan analisis diperdebatkan dan dikaji dalam diskusi. Kelebihan, kelemahan, serta implikasi praktis dari model Random Forest dalam prediksi stroke dibahas. Kesimpulan penelitian disusun berdasarkan temuan dan analisis yang dilakukan.

HASIL DAN PEMBAHASAN

Pada penelitian ini, dilakukan klasifikasi dan evaluasi performa model Random Forest untuk prediksi stroke. Random Forest merupakan salah satu metode machine learning yang efektif dalam melakukan klasifikasi pada dataset yang kompleks.

Dataset yang digunakan dalam penelitian ini adalah data stroke yang diperoleh dari sumber eksternal. Langkah pertama dalam pemrosesan data adalah membaca dataset menggunakan library pandas. Kemudian, dilakukan pengisian nilai yang hilang pada kolom bmi dengan menggunakan median dari data tersebut.

Keterangan Tabel :

A = id	G = work_type
B = gender	H = Residence_type
C = age	I = avg_glucose_level
D = hypertension	J = bmi
E = heart_disease	K = smoking_status
F = ever_married	L = stroke

Tabel 1 Data Stroke

A	B	C	D	E	F	G	H	I	J	K	L
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202,21	28,1	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	1

Selanjutnya, variabel kategorikal pada dataset seperti gender, ever_married, work_type, Residence_type, dan smoking_status diubah menjadi bentuk numerik menggunakan teknik LabelEncoder. Hal ini diperlukan karena model Random Forest hanya dapat memproses data numerik.

Tabel 2 Perubahan variabel kategorikal menjadi bentuk numerik

A	B	F	G	H	K
9046	1	1	1	1	1
51676	0	1	2	0	0
31112	1	1	1	0	0
60182	0	1	1	1	2

1665	0	1	2	0	0

Setelah itu, dilakukan pemisahan antara fitur dan target. Fitur-fitur yang digunakan untuk melakukan prediksi stroke disimpan dalam variabel X, sementara target (label) disimpan dalam variabel y. Kemudian, tipe data target juga diubah menjadi numerik menggunakan LabelEncoder

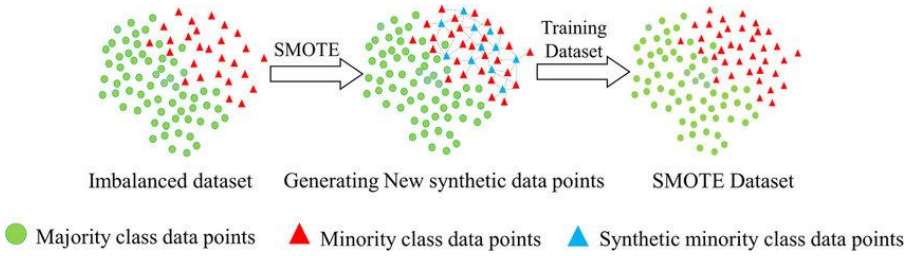
Tabel 3 Variabel X (Fitur)

A	B	C	D	E	F	G	H	I	J	K
9046	1	67	0	1	1	1	1	228,69	36,6	1
51676	0	61	0	0	1	2	0	202,21	28,1	0
31112	1	80	0	1	1	1	0	105,92	32,5	0
60182	0	49	0	0	1	1	1	171,23	34,4	2
1665	0	79	1	0	1	2	0	174,12	24	0

Tabel 4 Variabel Y (Target)

L
1
1
1
1
1

Untuk mengatasi ketidakseimbangan kelas pada dataset, dilakukan *oversampling* menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Hal ini bertujuan untuk menghasilkan dataset yang seimbang antara kelas mayoritas dan kelas minoritas.



Gambar 2 ***Synthetic Minority Over-sampling Technique***

Selanjutnya, data dibagi menjadi data pelatihan (training set) dan data pengujian (test set) menggunakan metode train_test_split dari library scikit-learn. Data pengujian sebesar 20% dari total data digunakan untuk mengevaluasi performa model yang telah dilatih.

Tabel 5 Data Pelatihan

A	B	C	D	E	F	G	H	I	J	K	L
51676	Female	61	0	0	Yes	Self-employed	Rural	202,21	28,1	never smoked	1

31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	1

Tabel 6 Data Pengujian

A	B	C	D	E	F	G	H	I	J	K	L
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	1

Setelah melakukan pelatihan model, dilakukan prediksi pada data pengujian. Hasil prediksi tersebut kemudian dievaluasi menggunakan beberapa metrik, yaitu akurasi (*accuracy*), presisi (*precision*), recall, dan F1-*Score*. Akurasi mengukur sejauh mana model dapat mengklasifikasikan data dengan benar secara keseluruhan, sedangkan presisi mengukur sejauh mana prediksi positif yang benar. *Recall* mengukur sejauh mana model dapat mendeteksi dengan benar kelas positif, sedangkan F1-*Score* merupakan perpaduan antara presisi dan recall.

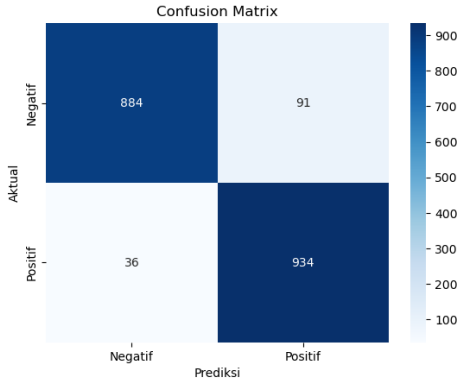
Akurasi :
Akurasi = (TP + TN) / (TP + TN + FP + FN)
= (933 + 887) / (933 + 887 + 88 + 37)
= 1820 / 1945 ≈ 0.936

Presisi:
Presisi = TP / (TP + FP)
= 933 / (933 + 88)
≈ 0.914

Recall:
Recall = TP / (TP + FN)
= 933 / (933 + 37)
≈ 0.961

F1-Score:
F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)
= 2 * (0.914 * 0.961) / (0.914 + 0.961)
≈ 0.937

Selain metrik-metrik tersebut, dilakukan juga perhitungan matriks kebingungan (*confusion matrix*) untuk memberikan gambaran lebih detail tentang performa model. *Confusion matrix* menampilkan jumlah prediksi yang benar dan salah untuk kelas positif dan kelas negatif.



Gambar 3 Hasil Evaluasi Confusion Matriks

Hasil evaluasi performa model Random Forest untuk prediksi stroke menunjukkan akurasi sebesar 93,6%, presisi sebesar 91,4%, *recall* sebesar 96,1%, dan F1-Score sebesar 93,7%. Confusion matrix menunjukkan bahwa terdapat 887 prediksi yang benar sebagai negatif (non-stroke) dan 933 prediksi yang benar sebagai positif (stroke). Sedangkan, terdapat 88 prediksi yang salah sebagai positif (false positive) dan 37 prediksi yang salah sebagai negatif (false negative).

Dengan hasil ini, dapat disimpulkan bahwa model Random Forest memiliki performa yang baik dalam melakukan prediksi stroke berdasarkan fitur-fitur yang digunakan dalam dataset.

SIMPULAN DAN SARAN

Kesimpulan

Dalam penelitian ini, telah dikembangkan dan dievaluasi model klasifikasi menggunakan algoritma Random Forest untuk prediksi stroke. Model ini dilatih menggunakan data klinis dan faktor risiko dari pasien yang terlibat dalam penelitian sebelumnya. Hasil evaluasi menunjukkan bahwa model Random Forest memiliki performa yang baik dengan tingkat akurasi, presisi, recall, dan skor F1 yang tinggi. Hal ini menunjukkan bahwa teknik machine learning, seperti Random Forest, dapat digunakan secara efektif dalam prediksi stroke berdasarkan data klinis dan faktor risiko.

Saran

Berdasarkan hasil penelitian ini, terdapat beberapa saran untuk pengembangan selanjutnya dalam bidang prediksi dan pencegahan stroke menggunakan machine learning:

1. Penelitian lanjutan: Dilakukan penelitian lebih lanjut untuk menjelajahi penggunaan model ensemble dan penggabungan data tambahan, seperti data genetik atau data lingkungan, guna meningkatkan akurasi prediksi stroke.
2. Optimisasi parameter: Melakukan optimisasi parameter model untuk meningkatkan kinerja dan akurasi prediksi stroke.
3. Pengujian validitas eksternal: Melakukan pengujian model menggunakan data eksternal yang tidak terlibat dalam penelitian awal untuk memastikan generalisasi yang lebih baik.
4. Studi lebih lanjut tentang faktor risiko: Dilakukan penelitian lebih lanjut tentang faktor risiko terkait stroke untuk memahami penyebab dan mekanisme penyakit ini secara lebih mendalam.

5. Penggunaan teknologi wearable: Menerapkan teknologi wearable dalam pemantauan pasien untuk mendeteksi dini stroke dan memberikan informasi yang berharga untuk prediksi dan penanganan.
6. Kolaborasi lintas-disiplin: Meningkatkan kolaborasi antara ilmu komputer, ilmu kesehatan, dan bidang terkait lainnya untuk menghasilkan solusi inovatif dalam prediksi dan pencegahan stroke.

DAFTAR PUSTAKA

Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C.-O., Chuah, J. H., & Gheni, H. M. (2022). Improving the heart disease detection and patients' survival using supervised infinite feature selection and improved weighted random forest. *IEEE Access*, *10*, 67363–67372.

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. https://doi.org/10.1613/jair.953

Coupland, A. P., Thapar, A., Qureshi, M. I., Jenkins, H., & Davies, A. H. (2017). The definition of stroke. *Journal of the Royal Society of Medicine*, *110*(1), 9–12. https://doi.org/10.1177/0141076816680121

dr. Rizal Fadli. (2022). *Stroke*. https://www.halodoc.com/kesehatan/stroke

Nedjar, I., Mahmoudi, S., & Chikh, M. A. (2022). A topological approach for mammographic density classification using a modified synthetic minority over-sampling technique algorithm. *International Journal of Biomedical Engineering and Technology*, *38*(2), 193–214.

Ramadhan, A., Susetyo, B., & Indahwati. (2019). PENERAPAN METODE KLASIFIKASI RANDOM FOREST DALAM MENGIDENTIFIKASI FAKTOR PENTING PENILAIAN MUTU PENDIDIKAN. *Jurnal Pendidikan Dan Kebudayaan*, *4*(2), 169–182. https://doi.org/10.24832/jpnk.v4i2.1327

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, *167*, 706–716. https://doi.org/10.1016/j.procs.2020.03.336

Wainberg, M., Alipanahi, B., & Frey, B. J. (2016). Are random forests truly the best classifiers? *The Journal of Machine Learning Research*, *17*(1), 3837–3841.

Wang, J., Yu, H., Hua, Q., Jing, S., Liu, Z., Peng, X., Cao, C., & Luo, Y. (2020). A descriptive study of random forest algorithm for predicting COVID-19 patients outcome. *PeerJ*, *8*, e9945. https://doi.org/10.7717/peerj.9945