Jurnal Ilmu Komputer dan Informatika (JIKI)
Vol. 3, No. 2, Desember 2023, Hal. 89-98
https://jiki.jurnal-id.com

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi *Decision Tree*, *Naïve Bayes*, dan *Random Forest*

Yunita Aulia[1], Andriyansyah[2], Suharjito[*3], Sri Wahyu Nensi[4]

[1,2,3,4]Industrial Engineering Department, BINUS Graduate Program – Master of Industrial Engineering, Bina Nusantara University, Jakarta, 11480, Indonesia.
Email: [1]yunita.aulia@binus.ac.id , [2]andriyansyah@binus.ac.id, [3]Suharjito@binus.edu, Sri.nensi@binus.ac.id

Abstrak

Prediksi *stroke* telah muncul sebagai bidang penelitian dan intervensi kesehatan yang penting karena dampaknya yang signifikan terhadap kesehatan masyarakat dan kesejahteraan individu. Pemeriksaan rinci mengenai usia, hipertensi, penyakit jantung, status perkawinan, jenis pekerjaan, jenis tempat tinggal, rata-rata kadar glukosa, BMI, status merokok, dan jenis kelamin sebagai factor terjadinya *stroke*. Dengan melakukan sintesis penelitian dan menganalisis kumpulan data yang luas, penelitian ini bertujuan untuk menjelaskan hubungan rumit antara faktor-faktor tersebut dan dampak kumulatifnya terhadap risiko *stroke*. Metode penelitian ini diawali dengan perbandingan algoritma *Decision Tree*, *Naïve Bayes* dan *Random Forest* dengan menggunakan *software RapidMiner*. Dari dataset prediksi *stroke* yang diberikan, terdapat 5110 responden dengan kondisi beragam. Di antara 5110 responden tersebut terdapat 12 atribut. Berdasarkan uraian yang telah dibahas maka dapat diambil kesimpulan bahwa metode *Decision Tree* merupakan metode terbaik dengan nilai akurasi tertinggi sebesar 95,13% dibandingkan dengan metode *Random Forest* dan *Naïve Bayes* dan nilai TF (*True False*) yang dipilih adalah 4861, TT (*True True*) adalah 0, FF (*False False*) adalah 249, dan FT (*False True*) adalah 0.

Kata kunci: *Decision Tree*, Klasifikasi, *Naïve Bayes*, Prediksi *Stroke*, *Random Forest*

Abstract

The prediction of stroke has emerged as a critical area of research and healthcare intervention due to its significant impact on public health and individual well-being. A detailed examination of age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and gender as predictors of stroke. By synthesizing existing research and analyzing extensive datasets, this study aims to elucidate the intricate relationships among these factors and their cumulative effect on stroke risk. These research methods are started by the comparison of Decision Tree, Naïve Bayes and Randon Forest algorithm by using RapidMiner software. From the given stroke prediction dataset, it comprises 5110 respondents with diverse conditions. Among these 5110 respondents, there are 12 attributes. Based on the discussed elaboration, the following conclusions can be drawn that decision tree method yields the highest accuracy value of 95.13% compared to the random forest and naïve bayes method and chosen TF (True False) value is 4861, TT (True True) is 0, FF (False False) is 249, and FT (False True) is 0.

Keywords: *Classification, Decision Tree, Naïve Bayes, Random Forest, Stroke Prediction*

1. PENDAHULUAN

Prediksi *stroke* telah muncul sebagai bidang penelitian dan intervensi kesehatan yang penting karena dampaknya yang signifikan terhadap kesehatan masyarakat dan kesejahteraan individu. *Stroke*, suatu kejadian *cerebrovascular* yang ditandai dengan gangguan tiba-tiba aliran darah ke otak, dapat mengakibatkan kerusakan neurologis yang parah dan seringkali tidak dapat diperbaiki lagi. Dalam beberapa tahun terakhir, penelitian ekstensif telah dilakukan untuk mengungkap hubungan rumit antara berbagai faktor risiko dan kemungkinan terjadinya *stroke*. Di antara faktor-faktor ini, usia, hipertensi, penyakit jantung, status perkawinan, jenis pekerjaan, jenis tempat tinggal, rata-rata kadar glukosa, BMI, status merokok, dan jenis kelamin muncul sebagai faktor penentu utama yang mempengaruhi kerentanan

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

stroke. Hubungan dan interaksi antara faktor-faktor ini sangat penting untuk prediksi, pencegahan, dan manajemen stroke yang efektif [1].

Usia, salah satu faktor risiko yang paling banyak diketahui, secara konsisten dikaitkan dengan peningkatan kejadian *stroke* [2]. Seiring bertambahnya usia seseorang, perubahan struktural dan fungsional pada pembuluh darah, serta efek kumulatif dari faktor risiko lainnya, berkontribusi terhadap peningkatan kerentanan terhadap *stroke*. Selain itu, hipertensi dan penyakit jantung, yang keduanya merupakan indikasi kesehatan kardiovaskular, memainkan peran penting dalam komplikasi vaskular yang menyebabkan stroke [3]. Hipertensi kronis dapat merusak pembuluh darah, berpotensi menyebabkan penggumpalan darah atau pecahnya pembuluh darah, yang keduanya dapat memicu *stroke*. Status perkawinan, jenis pekerjaan, dan jenis tempat tinggal mencerminkan pengaruh gaya hidup dan sosial ekonomi, yang dapat berdampak pada kesehatan seseorang secara keseluruhan dan risiko *stroke* [4].

Faktor metabolis seperti kadar glukosa rata-rata dan BMI memberikan wawasan tambahan mengenai prediksi *stroke*. Peningkatan kadar glukosa darah, yang merupakan ciri khas diabetes, berkontribusi terhadap peradangan dan kerusakan di dalam pembuluh darah, sehingga meningkatkan peluang risiko *stroke* [5]. Demikian pula, BMI yang lebih tinggi, yang mengindikasikan obesitas, dikaitkan dengan peningkatan kemungkinan berkembangnya faktor risiko lain seperti hipertensi dan diabetes. Merokok, merupakan salah satu faktor risiko lainnya, memiliki hubungan erat dengan kejadian *stroke* melalui efek buruknya pada pembuluh darah [6]. Merokok merusak pembuluh darah, mempercepat penumpukan plak arteri, dan meningkatkan kemungkinan penggumpalan darah, yang semuanya secara signifikan meningkatkan risiko *stroke*. Jenis kelamin juga terbukti mempengaruhi risiko *stroke*, dengan variasi profil risiko antara pria dan wanita [7]. Beberapa penelitian menunjukkan bahwa faktor risiko tertentu mungkin bermanifestasi secara berbeda pada pria dan Wanita.

Dalam bidang prediksi *stroke*, sejumlah besar penelitian terkait telah bermunculan, didorong oleh kebutuhan mendesak untuk memitigasi dampak buruk *stroke* terhadap kesehatan masyarakat. Sejumlah penelitian telah berupaya untuk mengidentifikasi prediksi yang dapat diandalkan dan mengembangkan model yang efektif untuk menilai risiko *stroke* individu. Investigasi ini mencakup beragam pendekatan, mulai dari metode statistik tradisional hingga teknik pembelajaran mesin yang lebih canggih. *Literatur* yang ada telah menyoroti pentingnya faktor risiko seperti hipertensi, diabetes, usia, dan pilihan gaya hidup yang berkontribusi terhadap terjadinya *stroke*. Para peneliti juga mengeksplorasi integrasi data pencitraan medis, penanda genetik, dan *biomarker* untuk meningkatkan akurasi prediksi. Dengan mengevaluasi secara kritis metodologi dan temuan penelitian sebelumnya, pemahaman komprehensif tentang penelitian prediksi *stroke* saat ini dapat dikembangkan, memberikan landasan yang kuat untuk pengembangan model prediksi yang inovatif dan kuat.

Dalam konteks penyakit *cerebrovaskular* yang parah seperti *stroke*, memprediksi kematian jangka pendek pada pasien merupakan hal yang penting secara medis. Penelitian ini menggunakan berbagai model pembelajaran mesin, termasuk Memanfaatkan pengklasifikasi *Random Forest* (RF), menggunakan *Adaptive Boosting* (AdaBoost), mengintegrasikan pengklasifikasi *Extremely Randomized Trees* (ExtraTree), memanfaatkan kekuatan pengklasifikasi *XGBoost*, memanfaatkan *TabNet*, dan menggabungkan *DistilBERT*, untuk membangun model prediksi yang menyeluruh. Model ini menggunakan data *bioassay* dan narasi teks radiologi dari pasien *stroke* hemoragik dan iskemik untuk memprediksi mortalitas dalam enam bulan. Langkah-langkah penilaian seperti area kurva karakteristik operasi penerima (AUROC), area kurva *recall presisi* (AUPRC), *presisi*, *recall*, dan skor F1 digunakan untuk mengukur keandalan model. Dengan mengintegrasikan data dari 19616 orang dengan *stroke* hemoragik dan 50178 orang dengan *stroke* iskemik, penelitian ini mengembangkan model prediksi baru untuk angka kematian dalam enam bulan. Model ini menunjukkan peningkatan kinerja dengan menggabungkan data uji laboratorium, informasi terstruktur, dan laporan radiologi tekstual. Hasil yang dicapai pasien hemoragik adalah: AUROC = 0.89, AUPRC = 0.70, *presisi* = 0.52, *recall* = 0.78, dan skor F1 = 0.63; sedangkan pada pasien iskemik diperoleh hasil: AUROC = 0,88, AUPRC = 0,54, *presisi* = 0,34, *recall* = 0,80, dan skor F1 = 0,48. Model prediksi ini mempunyai potensi untuk membantu dalam menilai risiko kematian dan mengidentifikasi pasien *stroke* risiko tinggi secara dini, sehingga menghasilkan alokasi sumber daya layanan kesehatan yang lebih efisien bagi para penyintas *stroke* [8].

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

Ringkasan yang diberikan menguraikan studi penelitian yang berfokus pada diagnosis dini *stroke* menggunakan sinyal elektrokardiogram (EKG). Penelitian bertujuan untuk membuat model klasifikasi yang efektif untuk mendiagnosis *stroke* menggunakan fitur EKG. Penelitian ini menggunakan data EKG dari 71 subjek, yang mencakup kohort 35 pasien *stroke* dan 36 orang tanpa kelainan medis. Pendekatan yang diusulkan melibatkan model ansambel bertumpuk, yang menggabungkan tiga model jaringan saraf konvolusional (CNN) yang berbeda. Sinyal EKG mentah digunakan sebagai masukan untuk pelatihan dan pengujian model. Hasilnya menunjukkan bahwa model yang dikembangkan mencapai akurasi luar biasa sebesar 99,7% dalam memprediksi *stroke*. Selain itu, metrik skor F1, presisi, dan perolehan juga sangat tinggi, masing-masing sebesar 99,69%, 99,67%, dan 99,71%. Kesimpulannya, penelitian ini menunjukkan bahwa model yang diusulkan menunjukkan potensi EKG sebagai alat yang efektif untuk membantu diagnosis *stroke* dengan tingkat efisiensi yang tinggi [9].

Beban *global* akibat *stroke* otak menggaris bawahi pentingnya intervensi yang tepat waktu, karena lebih dari 90% faktor risiko *stroke* dapat diantisipasi. Diagnosis yang cepat dan hemat biaya sangat penting dalam mengurangi komplikasi pasca *stroke*. Meskipun *Machine Learning* (ML) menawarkan cara untuk diagnosis dini, efektivitasnya berkurang dalam memprediksi kejadian langka dan menangani ketidakseimbangan kelas. Untuk mengatasi hal ini, penelitian ini memperkenalkan kerangka kerja ML baru untuk prediksi *stroke* otak, memanfaatkan *Artificial Immune Systems* (AIS) dan *Decision Trees* (DT) yang dikembangkan melalui *Genetic Programming* (GP). Berbeda dengan metode yang ada, pendekatan penelitian ini menekankan kemampuan interpretasi model, yang dicapai melalui operator penyederhanaan yang menyederhanakan DT yang diinduksi. Mengevaluasi kumpulan data yang tidak seimbang (1,89% kasus *stroke*), AIS dikombinasikan dengan *One Sided Selection* (OSS) memperbaiki ketidakseimbangan, memfasilitasi evolusi DT oleh dokter umum. Dalam eksperimen, pendekatan ini menghasilkan sensitivitas (70%) dan spesifisitas (78%), yang sebanding dengan teknik canggih. Selain itu, eksperimen kedua menunjukkan kapasitas metode dalam menghasilkan aturan yang dapat ditafsirkan manusia. Hasil ini menggaris bawahi potensi pendekatan penelitian untuk diagnosis *stroke* yang layak secara klinis, meningkatkan sensitivitas dan spesifisitas sambil mempertahankan interpretabilitas [10].

Studi ini menggunakan pendekatan pembelajaran mesin untuk meningkatkan akurasi diagnosis *stroke*, dengan fokus khusus pada mengatasi masalah ketidakseimbangan data. Keseimbangan dataset dicapai melalui pemanfaatan teknik *Random Over Sampling* (ROS). Investigasi ini menilai sebelas pengklasifikasi berbeda, yang mencakup metodologi seperti *Support Vector Machine*, *Random Forest*, *K-nearest Neighbor*, *Decision Tree*, *Naïve Bayes*, *Voting Classifier*, *AdaBoost*, *Gradient Boosting*, *Multi-Layer Perception*, dan *Nearest Centroid*. Awalnya, sebelum penyeimbangan kumpulan data, sepuluh klasifikasi menunjukkan tingkat akurasi melebihi 90%. Setelah penerapan *oversampling* untuk penyeimbangan kumpulan data, empat klasifikasi mencapai akurasi melebihi 96%. Penyetelan *hyperparameter* dan validasi silang dilakukan untuk mengoptimalkan *performa* model. Evaluasi *performa* model melibatkan berbagai metrik, termasuk Akurasi, Pengukuran F1, *Presisi*, dan *Recall*. Khususnya, *Support Vector Machine* muncul sebagai klasifikasi berkinerja tertinggi dengan akurasi 99,99%, disertai dengan nilai *recall*, *presisi*, dan ukuran F1 yang sama pada 99,99%. Klasifikasi *Random Forest* mencapai akurasi tertinggi kedua sebesar 99,87%, menunjukkan kesalahan minimal sebesar 0,001%. Lebih jauh lagi, penelitian ini memperluas pengaruhnya dengan mengembangkan aplikasi *web* dan *seluler* yang mudah digunakan berdasarkan model yang paling akurat. Perkembangan ini meningkatkan aksesibilitas dan kegunaan untuk prediksi *stroke*, sehingga memberikan kontribusi yang signifikan di bidang ini [11].

Kesenjangan penelitian dalam penelitian ini terhadap analisis prediksi *stroke* terletak pada evaluasi komparatif kemampuan prediksi *RapidMiner* terhadap teknik pembelajaran mesin konvensional lainnya. Meskipun *algoritma* yang ada telah digunakan secara luas untuk memperkirakan risiko *stroke*, keunggulan unik dari *RapidMiner* yang mudah digunakan dan alur kerja otomatis memperkenalkan kebutuhan untuk menilai efektivitasnya secara komprehensif dalam domain ini. Penelitian sebelumnya sebagian besar berfokus pada kinerja algoritmik dan akurasi prediksi, sering kali mengabaikan potensi intuitif *RapidMiner* untuk memberdayakan non-ahli dalam bidang perawatan kesehatan dan penelitian. Selain itu, terdapat kelangkaan penelitian yang menyelidiki kemampuan adaptasi *platform* terhadap

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

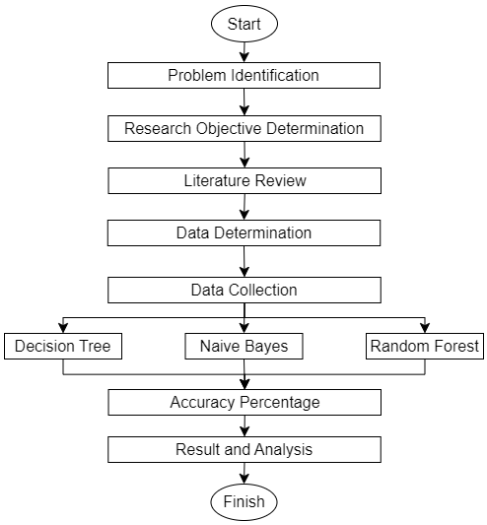
kumpulan data yang beragam, sejauh mana *platform* tersebut dapat mengakomodasi faktor risiko atau *biomarker* baru, dan potensinya untuk berkontribusi terhadap interpretasi dan penerapan klinis model prediksi *stroke*. Untuk menjembatani kesenjangan ini, eksplorasi komprehensif mengenai peran *RapidMiner* dalam prediksi *stroke*, termasuk perbandingan *head-to-head* dengan metodologi yang sudah ada, sangat penting untuk memberikan pemahaman yang lebih mendalam tentang kekuatan, keterbatasan, dan potensi kontribusinya untuk memajukan penilaian risiko *stroke*.

Mengingat hubungan yang rumit antara faktor-faktor ini, analisis yang komprehensif sangat penting untuk lebih memahami dampak kolektifnya terhadap prediksi *stroke*. Teknik analisis data tingkat lanjut, termasuk pembelajaran mesin dan pemodelan statistik, memungkinkan eksplorasi lebih dalam mengenai sifat multidimensi risiko *stroke*. Wawasan yang diperoleh dari analisis tersebut berpotensi untuk menyempurnakan alat penilaian risiko dan memandu intervensi yang ditargetkan, yang pada akhirnya mengurangi beban *stroke* pada individu dan sistem layanan kesehatan.

Terkait pemanfaatan teknologi, melakukan penelitian menggunakan *RapidMiner* sebagai perangkat lunak pembantu pemrosesan data menawarkan pendekatan analisis data yang efisien. *RapidMiner* yang yang mudah digunakan memudahkan para peneliti, terlepas dari keahlian teknis mereka, untuk dengan mudah melakukan praproses, membersihkan, dan memanipulasi kumpulan data yang kompleks. Transformasi data yang beragam dan alat rekayasa fitur memungkinkan persiapan data yang komprehensif, memastikan bahwa penelitian dibangun di atas dasar yang kuat. Selain itu, kemampuan pembelajaran mesin dan pemodelan prediktif *RapidMiner* yang kuat memungkinkan para peneliti memperoleh wawasan yang bermakna dari data, membantu dalam identifikasi pola, tren, dan prediksi potensial yang terkait dengan fokus penelitian, seperti prediksi *stroke*. Alat visualisasi *platform* memfasilitasi komunikasi temuan yang jelas, sehingga meningkatkan dampak dan penerapan penelitian. Dengan memanfaatkan kemampuan *RapidMiner*, para peneliti dapat mempercepat proses penelitian, membuat keputusan yang tepat, dan berkontribusi terhadap kemajuan dalam prediksi *stroke* dengan ketelitian.

Dalam tulisan ini, kami menyajikan pemeriksaan rinci mengenai usia, hipertensi, penyakit jantung, status perkawinan, jenis pekerjaan, jenis tempat tinggal, rata-rata kadar glukosa, BMI, status merokok, dan jenis kelamin sebagai prediksi *stroke*. Dengan melakukan sintesis penelitian yang ada dan menganalisis kumpulan data yang luas, kami bertujuan untuk menjelaskan hubungan rumit antara faktor-faktor ini dan dampak kumulatifnya terhadap risiko *stroke*. Hasil penelitian ini berkontribusi pada pemahaman yang lebih holistik mengenai prediksi *stroke* dan menawarkan wawasan berharga bagi praktisi kesehatan, pembuat kebijakan, dan peneliti yang bekerja menuju strategi pencegahan *stroke* yang efektif.

2. METODE PENELITIAN



Gambar 1. *Research Flow*

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

Metode penelitian ini diawali dengan perbandingan *algoritma Decision Tree*, *Naïve Bayes* dan *Random Forest* untuk analisis prediksi *stroke*. Data dikumpulkan dari situs online seperti kaggle.com. Setelah data dipilih, data perlu dilanjutkan ke tahap *pra*-pemrosesan. Selanjutnya *algoritma Decision Tree, Naïve Bayes* dan *Random Forest* diterapkan pada data yang siap diolah melalui beberapa kali uji coba. Penelitian dilakukan dengan menggunakan persentase akurasi.

Decision trees mewakili teknik ampuh yang sering digunakan di berbagai domain, termasuk pembelajaran mesin, pemrosesan gambar, dan pengenalan pola [12]. Dalam membangun model prediktif, *Algoritma DT-Quest* 'diselaraskan' dengan parameter proses eksperimen. Hal ini melibatkan pemanfaatan bobot *Decision trees* untuk memilih latihan secara dinamis, memfasilitasi evaluasi prediksi *stroke* [13].

Untuk selanjutnya, deteksi bug perangkat lunak sebelum peluncuran terjadi melalui beberapa tahap dan mendapat prioritas yang signifikan. Antisipasi terhadap kerusakan perangkat lunak mendapat perhatian penelitian yang luas karena hal ini sangat penting dalam sektor perangkat lunak. *Naive Bayes* menonjol sebagai pendekatan pembelajaran yang dominan (merupakan 47,4% dari seluruh investigasi) dalam bidang prediksi kerusakan perangkat lunak [14]. NB (*Naive Bayes*) menggunakan teorema *Bayes* dengan cara sebagai berikut:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (1)$$

Catatan;
Y = variabel kelas
X = vektor fitur bergantung n-dimensi

Kesalahan *random forest* diperkirakan menggunakan kesalahan *out-of-bag* (OOB) selama fase pelatihan. Setiap pohon dibuat dari sampel *bootstrap* yang berbeda, dengan sekitar sepertiga observasi dihilangkan secara acak. Pengamatan yang dikecualikan ini, membentuk sampel OOB untuk pohon tertentu, sangat penting dalam mengukur kinerja model. Patut dicatat bahwa dalam *algoritme random forest*, ukuran subset variabel prediktor (dilambangkan dengan "m") memainkan peran penting dalam mengendalikan kedalaman pohon. Oleh karena itu, menyempurnakan parameter ini selama pemilihan model sangatlah penting, subjek yang akan diuraikan dalam contoh berikutnya [16].
Jika nilainya menunjukkan kedekatan, himpunan tersebut dapat dianggap tepat. Ketika *mean* hampir sejajar dengan nilai sebenarnya dari besaran yang diukur, himpunan tersebut mencapai akurasi [17]. Istilah-istilah ini hanya dapat dievaluasi jika dilengkapi dengan kumpulan data pengukuran berulang untuk kuantitas yang sama.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

Catatan:
TP = *True positive*
TN = *True negative*
FP = *False positive*
FN = *False negative*

3. HASIL DAN PEMBAHASAN

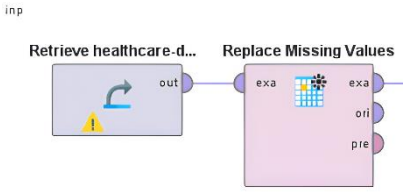
Dari dataset prediksi *stroke* yang diberikan, terdapat 5110 responden dengan kondisi beragam. Di antara 5110 responden tersebut terdapat 12 atribut yaitu: jenis kelamin, umur, hipertensi, penyakit jantung, pernah menikah, jenis pekerjaan, tipe tempat tinggal, kadar glukosa rata-rata, BMI, status merokok, dan *stroke*. Dari 12 atribut tersebut, atribut “id” dikecualikan karena tidak mempengaruhi hasil.

Ketiga model tersebut akan diuji dan dievaluasi menggunakan *software Rapid Miner* untuk mendapatkan hasil yang paling akurat. Seluruh atribut tersebut akan dijadikan ukuran untuk memprediksi kondisi *stroke* responden.

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

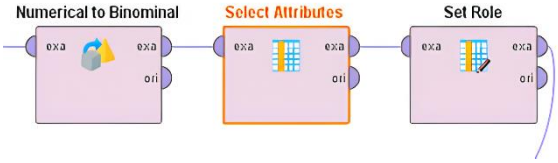
Untuk mencapai akurasi maksimum, pra-pemrosesan dilakukan terlebih dahulu. Pra-pemrosesan digunakan untuk menghilangkan nilai yang hilang, *outlier*, pengkodean label, dan lain sebagainya, dengan tujuan untuk mendapatkan hasil yang optimal.

3.1. *RapidMiner Design*



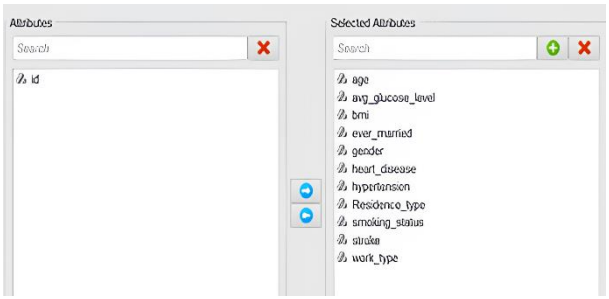
Gambar 2. *Using Operator Pre-processing*

Gambar 2 menunjukkan awal pembentukan model dengan melakukan *pre-processing* data menggunakan *replace missing values.*



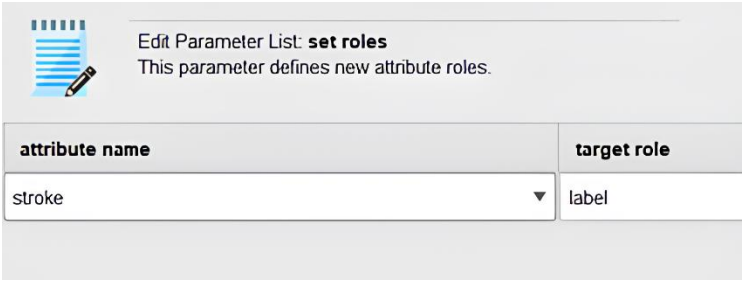
Gambar 3. *Required Data Binomial*

Gambar 3 menunjukkan hubungan antar data yang diperlukan dalam proses klasifikasi 3 metode dalam bentuk binomial.



Gambar 4. *Selection Attributes.*

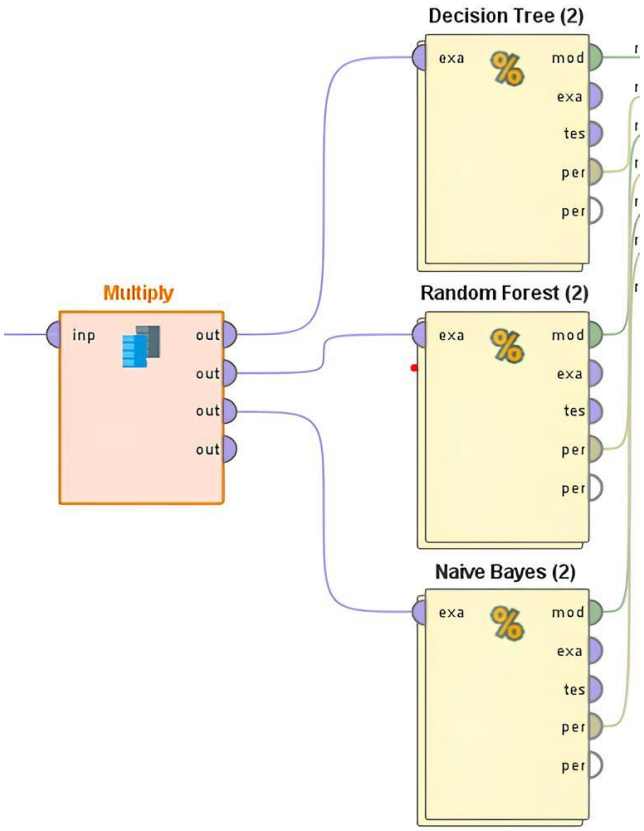
Gambar 4 menggambarkan pemilihan 11 atribut sebagai data yang diolah. Kolom “id” tidak dipilih sebagai atribut karena tidak mempengaruhi hasil pengolahan data.



Gambar 5. *Detail Parameter Roles*

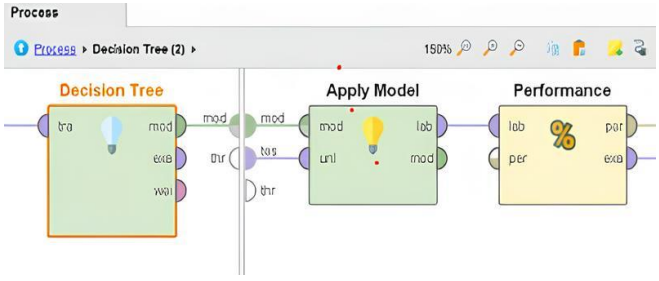
Gambar 5 menunjukkan kolom “*Stroke*” sebagai label klasifikasi.

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

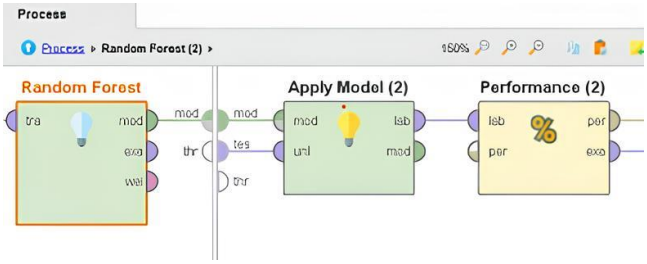


Gambar 6. *Detail Combine Three Method*

Gambar 6, operator "*Multiply*" diperlukan untuk menggabungkan beberapa model menjadi satu desain *Rapid Miner*. Operator "*Multiply*" ini terhubung ke model *decision tree*, *random forest*, dan *naïve bayes* menggunakan operator "*cross validation*".

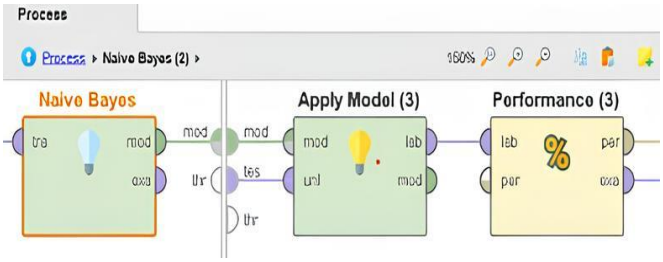


Gambar 7. *Detail Operator Decision Tree*



Gambar 8. *Detail Operator Random Forest*

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591



Gambar 9. *Detail Operator Naïve Bayes*

Masing-masing operator "*cross validation*" akan membentuk desain seperti gambar di atas: *decision tree*, pada Gambar 7, *random forest* pada Gambar 8, dan *naïve bayes* pada Gambar 9. Keseluruhan proses akan dijalankan untuk mendapatkan hasil kinerja klasifikasi dari ketiga model ini .

3.2. Klasifikasi Hasil

accuracy : 95.13% +/-0.06% (micro average: 95.13%)

	true false	true true	class precision
pred. false	4861	249	95.13%
pred. true	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 10. *Result Classification Decision Tree*

Gambar 10 menggambarkan hasil pengujian akurasi model *decision tree*,. Dapat diamati nilai akurasinya sebesar 95,13%. Penarikan kembali kelas untuk benar-salah adalah 100% dan benar-benar adalah 0%. Sedangkan *presisi* kelas untuk *false* sebesar 95,13% dan untuk *true* sebesar 0%.

accuracy : 95.03% +/-0.25% (micro average: 95.03%)

	true false	true true	class precision
pred. false	4847	240	95.28%
pred. true	0	0	0.00%
class recall	99.71%	3.61%	

Gambar 11. *Result Classification Random Forest*

Gambar 11 menggambarkan hasil pengujian akurasi model *random forest*. Dapat diamati nilai akurasinya sebesar 95,03%. *Recall kelas* untuk benar-salah sebesar 99,71% dan benar-benar sebesar 3,61%. Sedangkan *presisi* kelas untuk *false* sebesar 95,28% dan *true* sebesar 39,13%.

accuracy : 92.92% +/-0.90% (micro average: 92.92%)

	true false	true true	class precision
pred. false	4738	236	95.20%
pred. true	123	10	7.52%
class recall	97.47%	4.02%	

Gambar 12. *Result Classification Naïve Bayes*

Gambar 12 menyajikan hasil pengujian akurasi model *naïve bayes*. Dapat diamati nilai akurasinya sebesar 92,92%. Recall kelas untuk benar-salah sebesar 97,47% dan benar-benar sebesar 4,02%. Sedangkan *presisi* kelas untuk *false* sebesar 95,20% dan untuk *true* sebesar 7,52%.

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

METHOD	ACCURACY (%)
Decision Tree	95.13%
Random Forest	95.03%
Naïve Bayes	92.92%

Gambar 13. *Comparation Accuracy Three Method*

Pada Gambar 13 di atas terlihat bahwa metode *decision tree* memberikan nilai akurasi sebesar 95,13%, lebih tinggi dibandingkan dengan *random forest* dan *naïve bayes*.

4. KESIMPULAN

Berdasarkan uraian yang telah dibahas, dapat diambil kesimpulan sebagai berikut: Metode *decision tree* terpilih sebagai metode terbaik dengan nilai akurasi tertinggi sebesar 95,13% sedangkan metode *random forest* dan *naïve bayes* dan nilai TF (*True False*) yang dipilih adalah 4861, TT (*True True*) adalah 0, FF (*False False*) adalah 249, dan FT (*False True*) adalah 0. *Stroke* merupakan salah satu ancaman kesehatan berbahaya di seluruh dunia. Bagi mereka yang mengalami gejala atau pernah menderita *stroke*, pemulihan segera sangat penting untuk mencegah kerusakan lebih lanjut. Dalam hal ini, model pembelajaran mesin dapat berperan penting dalam memprediksi gejala awal *stroke*. Makalah ini mendeteksi dan memprediksi hasil stroke berdasarkan 11 klasifikasi menggunakan berbagai metode pembelajaran mesin.

Penelitian di masa depan akan menggunakan metode pembelajaran mendalam untuk mencapai hasil yang lebih akurat. Dengan penelitian yang sedang berlangsung, kami optimis dapat mengurangi prevalensi penyakit ini dengan cepat.

DAFTAR PUSTAKA

[1] J. Li, Y. Luo, M. Dong, Y. Liang, X. Zhao, Y. Zhang and Z. Ge, "Tree-Based Risk Factor Identification and Stroke Level Prediction in Stroke Cohort Study," *Biomed Research International*, pp. 1-10, 2023.

[2] R. D. Nindrea and A. Hasanuddin, "Non-modifiable and modifiable factors contributing to recurrent stroke: A systematic review and meta-analysis," *Clinical Epidemiology and Global Health* , p. 101240, 2023.

[3] F. D. Fuchs and P. K. Whelton, "High Blood Pressure and Cardiovascular Disease," *Hypertension*, vol. 75, no. 2, pp. 285-292, 2020.

[4] D. Puciato, M. Rozpara, M. Bugdol and B. Mróz-Gorgoń , "Socio-economic correlates of quality of life in single and married urban individuals: a Polish case study," *Health Qual Life Outcomes*, vol. 20, no. 58, 2022.

[5] M. Bakhtiyari, E. Kazemian, K. Kabir, F. Hadaegh, S. Aghajanian, P. Mardi, N. T. Ghahfarokhi, A. Ghanbari, M. A. Mansournia and F. Azizi, "Contribution of obesity and cardiometabolic risk factors in developing cardiovascular disease: a population-based cohort study," *Scientific Reports*, vol. 12, 2022.

[6] J. Chen, S. Li, K. Zheng, H. Wang, Y. Xie, P. Xu, Z. Dai, M. Gu, Y. Xia, M. Zhao, X. Liu and G. Xu, "Impact of Smoking Status on Stroke Recurrence," *Journal of the American Heart Association*, 2019.

[7] C. W. Yoon and C. D. Bushnell, "Stroke in Women: A Review Focused on Epidemiology, Risk Factors, and Outcomes," *Journal of Stroke*, vol. 25, no. 1, pp. 2-15, 2023.

[8] R. Huang, J. Liu, T. K. Wan, D. Siriwanna, Y. M. P. Woo, A. Vodencarevic, C. W. Wong and K. H. K. Chan, "Stroke mortality prediction based on ensemble learning and the combination of structured and textual data," Computers in Biology and Medicine, p. 106176, 2023.

DOI: https://doi.org/10.54082/jiki.90
P-ISSN: 2807-6664
E-ISSN: 2807-6591

[9] P. Kunwar and P. Choudhary, "A stacked ensemble model for automatic stroke prediction using only raw electrocardiogram," *Intelligent Systems with Applications,* p. 200165, 2023.

[10] L. I. Santos, M. O. Camargos, M. F. S. V. D'Angelo, J. B. Mendes, E. E. C. d. Medeiros, A. L. S. Guimarães and R. M. Palhares, "Decision tree and artificial immune systems for stroke prediction in imbalanced data," *Expert Systems with Applications,* vol. 191, 2022.

[11] N. Biswas, K. M. M. Uddin, S. T. Rikta and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics,* vol. 2, 2022.

[12] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends,* pp. 20-28, 2021.

[13] V. Matzavela and E. Alepis, "Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments," *Computers and Education: Artificial Intelligence,* p. 100035, 2021

[14] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing,* 2020.

[15] V. Sheth, U. Tripathi and A. Sharma, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," *Procedia Computer Science 215,* p. 422–431, 2022.

[16] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal,* pp. 3-29, 2020.

[17] . H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *International Journal of Computer Sciences and Engineering,* pp. 74-78, 2018.