Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

Analisis Perbandingan Algoritma Naïve Bayes dan Random Forest Dalam Klasifikasi Penyakit Stroke Pada Puskesmas

Iwan Virgiawan, Erizal*

Fakultas Teknologi Industri dan Informatika, Program Studi Teknik Informatika, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia
Email: [1]iwanvi02@gmail.com, [2,*]erizal@uhamka.ac.id
Email Penulis Korespondensi: erizal@uhamka.ac.id
Submitted: **17/01/2025**; Accepted: **28/03/2025**; Published**: 28/03/2025**

Abstrak−Salah satu alasan utama orang menjadi cacat atau meninggal adalah karena stroke. Kunci untuk terapi yang cepat dan efektif adalah diagnosis dini. Penelitian ini mengkaji kinerja relatif algoritma Naïve Bayes dan Random Forest dalam mengidentifikasi kasus stroke menggunakan data yang dikumpulkan dari pasien di Puskesmas Cipayung. Usia, jenis kelamin, BMI, status merokok, hipertensi, dan masalah kesehatan fisik dan mental lainnya adalah beberapa karakteristik yang terwakili dalam 644 sampel yang digunakan dalam penelitian ini. Mengumpulkan data, membersihkannya, dan kemudian mengevaluasi model menggunakan metrik seperti mengingat kembali, presisi, dan akurasi adalah bagian dari proses penelitian. Dengan tingkat akurasi precision 92%, Recall 92% dan f1-score 88% untuk random forest dalam mengklasifikasi penyakit stroke, sedangkan metode naïve bayes mencapai tingkat keberhasilan score precision 89%, Recall 87% dan f1-score 88% menurut data tersebut. Profesional medis dapat menggunakan hasil ini sebagai panduan untuk meningkatkan deteksi stroke, yang pada gilirannya mempercepat perawatan dan mengurangi kemungkinan konsekuensi. Temuan penelitian ini juga membuka jalan bagi penelitian masa depan tentang algoritma pembelajaran mesin.

Kata Kunci: Stroke; Klasifikasi; Naïve Bayes; Random Forest; Machine Learning

Abstract−One of the main reasons people become disabled or die is because of a stroke. The key to swift and effective therapy is an early diagnosis. This research examines the relative performance of the Naïve Bayes and Random Forest algorithms in identifying stroke cases using data collected from patients at the Cipayung Health Center. Age, gender, BMI, smoking status, hypertension, and other physical and mental health issues are some of the characteristics represented in the 644 samples used in the study. Collecting data, cleaning it up, and then evaluating the model using metrics like recall, precision, and accuracy are all part of the research process. With a 92% accuracy rate, the Random Forest algorithm outperformed Naïve Bayes (87% accuracy rate), according to the data. Medical professionals may use these results as a guide to improve stroke detection, which in turn accelerates treatment and lessens the likelihood of consequences. The findings of this study also pave the way for future research into machine learning algorithms.

Keywords: Stroke; Classification; Naïve Bayes; Random Forest; Machine Learning

1. PENDAHULUAN

Teknologi di era informasi telah berkembang seiring dengan periode modern, memfasilitasi penyelesaian sejumlah besar masalah yang lebih rumit. Salah satunya adalah serangkaian langkah menggunakan metode statistik; ini ada hubungannya dengan penambangan data. Komputer dapat menganalisis data tertentu untuk menghasilkan model yang dapat digunakan tanpa memasukkan kode program berulang kali ke dalam proses input-output; ini memungkinkan ekstraksi dan identifikasi informasi dan pengetahuan berdasarkan data besar di bidang-bidang seperti matematika, kecerdasan buatan, dan pembelajaran mesin [1]. Pembelajaran terbimbing dan tak terbimbing adalah dua pendekatan utama pembelajaran mesin [2]. Ketika hasil yang diinginkan atau diharapkan diketahui sebelumnya, pembelajaran terbimbing menjadi alat yang efektif. Klasifikasi termasuk dalam kelompok pembelajaran terbimbing [3]. Definisi tambahan menyatakan bahwa item data dievaluasi selama kategorisasi untuk ditempatkan ke dalam salah satu dari beberapa kelas yang sudah ada sebelumnya. Tahap pertama klasifikasi adalah membangun model untuk disimpan dalam memori. Langkah kedua adalah menerapkan model ke item data tambahan untuk menentukan kelasnya dalam model yang tersimpan [4].

Satu dari setiap lima kematian dan kecacatan secara global disebabkan oleh stroke. Pencitraan otak dan riwayat pasien membantu dokter membedakan antara stroke iskemik dan pendarahan intraserebral [5]. Dalam skala global, stroke menempati peringkat tinggi di antara penyebab utama kecacatan. Setiap tahun, 15 juta orang mengalami stroke; di antara mereka, 13% meninggal karena penyakit tersebut dan 5% sisanya mengalami gangguan jangka panjang. Setelah stroke, penyakit jantung iskemik merenggut nyawa lebih banyak orang daripada penyebab tunggal lainnya. Stroke semakin umum terjadi di Indonesia, yang memiliki jumlah korban stroke terbanyak di Asia, menurut statistik Riskesdas tahun 2018. Stroke mempengaruhi 13,7% populasi di kawasan Asia-Pasifik pada tahun 2005, dan para ahli memperkirakan angka itu akan meningkat menjadi 64,6% juta pada tahun 2050 [6].

Pada dasarnya puskesmas cipayung masih tergolong cukup tertinggal dalam hal Identifikasi dan klasifikasi faktor risiko stroke menghadirkan rintangan substansial bagi Puskesmas Cipayung. Metode teknologi, seperti algoritma pembelajaran mesin, mungkin memegang kunci untuk memecahkan masalah yang sulit ini. Ketika mencoba untuk memprediksi seberapa besar kemungkinan seseorang akan terkena stroke, kategorisasi stroke merupakan langkah penting. Tenaga medis pada Puskesmas Cipayung dapat mendeteksi kejadian stroke dengan lebih baik jika mereka memahami tanda-tanda dan faktor risiko penyakit kardiovaskularp. Indikasi peringatan dini masalah

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

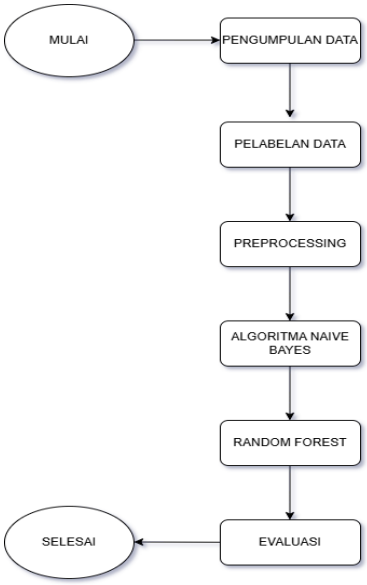
kesehatan, seperti bahaya stroke, juga dapat dideteksi dengan penggunaan teknologi informasi, khususnya pembelajaran mesin. Pekerjaan ini menggunakan dua algoritma pembelajaran mesin, Random Forest dan Naive Bayes, yang telah menunjukkan kemanjuran dalam klasifikasi penyakit menggunakan data klinis. Teknologi baru membantu para profesional medis dalam kategorisasi kasus stroke. Sebagai pendekatan prediksi dalam klasifikasi, ia mencari model dan fungsi yang dapat menjelaskan atau membedakan kelas data untuk memperkirakan kelas objek yang labelnya tidak tersedia [7]. Dengan menggunakan nilai probabilitas yang diberikan oleh setiap kriteria, kedua algoritma ini dapat membantu dalam klasifikasi dan analisis penyakit stroke. Dalam pemilihan fitur, pemilihan maju adalah teknik umum untuk menemukan nilai terbaik. Performa sistem klasifikasi dievaluasi menggunakan F-Measure, yang menggabungkan Precision dan Recall; nilai yang lebih besar menunjukkan performa yang lebih baik [8]. Naïve Bayes Classifier, sering dikenal sebagai teknik Naïve Bayes, adalah salah satu pendekatan untuk menggunakan data pengalaman masa lalu untuk memperkirakan kemungkinan potensial dan memastikan nilai probabilitasnya [9]. Algoritma Random Forest didasarkan pada metodologi pohon keputusan yang dapat beradaptasi. Pendekatan ini dapat mengungguli Naive Bayes dalam hal akurasi klasifikasi karena penggunaan beberapa pohon, yang menggabungkan output dari banyak pohon keputusan [10]. Breiman pertama kali menyarankan teknik Random forest untuk membangun prediktor menggunakan kumpulan pohon keputusan yang dihasilkan secara acak di dalam subruang data. kesebelas Setiap pertanyaan dalam pohon keputusan mungkin memiliki cabang tergantung pada nilai atribut dan akhirnya berakhir pada daun, yang memberikan estimasi untuk variabel kelas l. Pertanyaan-pertanyaan tersebut disusun secara metodis [11]. Di antara banyak manfaat pendekatan hutan acak adalah tingkat kesalahannya yang rendah, kinerja klasifikasi yang tinggi, penanganan cepat kumpulan data pelatihan besar-besaran, dan kemampuan untuk memperkirakan data yang hilang secara efektif [12].

Penelitian terdahulu yang dilakukan oleh Fitri Adha dengan judul "Perbandingan Metode Data Mining untuk Klasifikasi Penyakit Stroke" oleh Tati Suprapti, Hariyati Airi, dan Agus Bahtiar menunjukkan bahwa algoritma Random Forest lebih unggul dibandingkan dengan teknik Naive Bayes dalam upaya penelitian ini untuk mengklasifikasi penyakit stroke. Penggunaan metode Naive Bayes dengan 90% data latih dan 10% data uji menghasilkan nilai akurasi sebesar 71,9%, presisi sebesar 71,1%, dan recall sebesar 71,9%. Selain itu, Random Forest memperoleh nilai recall, presisi, dan akurasi sebesar 92,5 persen [13]. Berdasarkan uraian di atas, maka akan dilakukan perbandingan antara pendekatan Naive Bayes dengan pendekatan random forest untuk klasifikasi penyakit stroke. Sumber data yang kami gunakan adalah Puskesmas Cipayung. Karena deteksi dini memungkinkan penanganan yang lebih cepat, penulis penelitian ini berharap agar hasil penelitian ini dapat mempermudah tenaga medis dalam mengidentifikasi penyakit stroke.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Berikut Gambar 1 merupakan tahapan dalam melakukan penelitian.



Gambar 1. Tahapan Penelitian

Dari tahapan Gambar 1 dapat dijelaskan:

2.1.1 Pengumpulan Data

Pendekatan pengumpulan data terbaik digunakan oleh penulis. Sebagai teknik pengumpulan data, Premier Data memperoleh temuannya langsung dari sumbernya. Fasilitas kesehatan Cipayung merupakan sumber data yang

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

digunakan oleh penulis. Meskipun hasil atau data keluaran hanya digunakan untuk tujuan pengujian, data ini juga dapat digunakan sebagai data pelatihan. Selama dua belas bulan terakhir, penulis telah menggunakan 644 data. Berikut ini adalah garis besar karakteristik terkait stroke yang ditemukan dalam kumpulan data:

Tabel 1. Atribut pada column

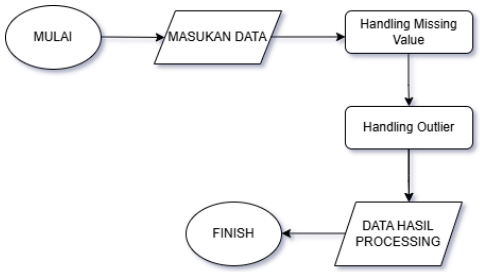
Atribut	Keterangan
Sex	Jenis Kelamin Pasien
Age	Umur Pasien
Chol check	kadar kolesterol dalam darah
Heart DiseaseorAttack	riwayat penyakit jantung atau serangan jantung.
High chol	Tingginya kadar kolesterol dalam darah
PhysActivity	Aktivitas fisik, yang mencakup semua bentuk gerakan tubuh
HvyAlcoholConsump	Konsumsi alkohol berat
BMI	Indeks Massa Tubuh
Smoking_status	Merokok
GenHlth	mencerminkan persepsi individu tentang kesehatan umum seperti "sangat baik", "baik", "cukup", "buruk", atau "sangat buruk".
Fruits	Mengacu pada frekuensi atau jumlah konsumsi buah dalam pola makan
Veggies	Mengacu pada konsumsi sayuran dalam pola makan
DiffWalk	merujuk pada kesulitan seseorang untuk berjalan atau bergerak
HighBP	tekanan darah tinggi atau hipertensi
MentHlth	mencerminkan kondisi kesehatan mental
PhysHlth	menggambarkan kondisi kesehatan fisik secara keseluruhan
Stroke	Apakah pasien mengalami stroke
Diabetes	Penyakit kronis yang ditandai dengan kadar gula darah (glukosa) yang tinggi

2.2.2 Pelabelan Data

Data yang terkumpul kemudian dipindahkan ke tahap pelabelan data. Potensial dan non-potensial merupakan dua (2) perasaan yang harus ditemukan dalam data pelabelan data dilakukan berdasarkan dari tabel yang disudah dijelaskan pada tabel 1 [14]. Saya mengolah data secara manual ke dalam file CSV menggunakan Microsoft Excel setelah memberi label pada setiap bagian data untuk pemeriksaan.

2.2.3 Preprocessing

Sebelum digunakan oleh sistem, data asli harus melalui pra-pemrosesan. Oleh karena itu, untuk meningkatkan kualitas beberapa data, perlu dilakukan sejumlah proses pra-pemrosesan. Data harus dibersihkan pada tahap pra-pemrosesan sebelum langkah pemodelan dapat dilaksanakan dalam penelitian ini.



Gambar 2. Flowchart Preprocessing

Pada awal fase praproses, data masukan disiapkan untuk menangani nilai yang hilang. Kesalahan entri data atau tidak ada data sama sekali dapat menyebabkan angka hilang. Oleh karena itu, variabel yang hilang perlu diselesaikan sebelum memulai model. Pendekatan imputasi rata-rata digunakan oleh peneliti. Untuk melengkapi data yang hilang, nilai rata-rata kolom harus digunakan [15].

Setelah itu, kita akan menggunakan pendekatan Z-score untuk menangani outlier. Tujuan dari prosedur ini adalah untuk mengetahui apakah data mengandung outlier nilai yang sangat ekstrem. Ketika Z-score kurang dari -3 atau lebih dari +3, sering kali berarti bahwa data tersebut mengandung nilai ekstrem. Jika ukuran data melebihi atau melampaui batas ini, data tersebut akan dihapus [16]. Tidak perlu menormalkan data untuk mendeteksi outlier karena data yang diperoleh bersih.

2.2.4 Algoritma Naive Bayes

Naive Bayes adalah pengklasifikasi probabilistik langsung yang menggunakan kumpulan frekuensi data dan kombinasi nilai untuk menghasilkan sekumpulan probabilitas. Prinsip dasar teorema Bayes adalah bahwa atribut tidak boleh bergantung pada atau dipengaruhi oleh nilai yang diberikan ke variabel kelas. Selama Anda memiliki data

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

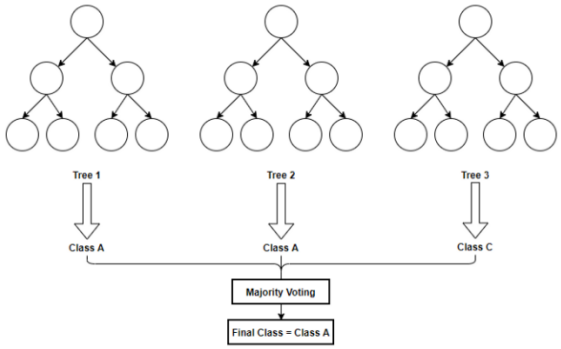
pelatihan dalam jumlah yang sedikit, pendekatan ini dapat menemukan estimasi parameter proses klasifikasi [17].

$$P(c|x) = \frac{P(x|c).P(c)}{P(x)} \quad (1)$$

Rumus Naive Bayes P(c|x) adalah Posterior | probability yaitu nilai probabilitas x berdasarkan kondisi c, P(c|x) adalah Probabilitas c yang ditentukan x adalah benar, P(x) adalah Peluang evidence penyakit x, dan P(c) adalah Probabilitas dari nilai c.

2.2.5 Random Forest

Tujuan utama dari penelitian ini adalah untuk menggunakan model kategorisasi *Random Forest*. Karena metode ini didasarkan pada pohon keputusan, sejumlah besar pohon keputusan akan dihasilkan selama pelatihan *Random Forest* menggunakan sampel set pelatihan [18].



Gambar 3. Algoritma *Random Forest*

Hutan yang Tidak Terkendali mengharuskan penggunaan banyak pohon keputusan untuk memperkirakan hasil secara akurat. Hasil yang dihasilkan setiap pohon keputusan saat menggunakan hutan acak sebagai pengklasifikasi mungkin sama atau berbeda. Misalnya, pohon keputusan C dan D memperkirakan 0 hasil, tetapi pohon keputusan A, B, E, dan F memperkirakan 1. Hutan acak memilih kesimpulan yang diproyeksikan karena ada banyak respons potensial dan kemungkinan yang tinggi. Kombinasi suara mayoritas dan hasil yang diharapkan dari pohon keputusan lainnya menentukan hasil dari beberapa pohon keputusan.

2.2.6 Evaluasi

Tujuan dari melakukan penilaian adalah untuk memastikan bahwa ujian tersebut akurat. Pengujian dilakukan untuk mendapatkan hasil terbaik dan untuk mengukur seberapa akurat temuan tersebut. Dengan membandingkan akurasi, presisi, dan perolehan kembali model, kita dapat melihat seberapa baik matriks kebingungan merepresentasikan kualitas model [19].

Tabel 2. Confusion Matrix

	Positif	Negatif
Positif	True Positif (TP)	False Negatif (FN)
Negatif	True Negatif (TN)	False Positif (FP)

Di Tabel 2 merupakan confusion matrix yang berisi True Positif (TP), True Negatif (TN), False Positif (FP) dan False Negatif (FN). Presisi merupakan bagian dari akurasi yang berfokus pada kebenaran prediksi model, sedangkan akurasi menguji seberapa berhasil model dapat mengklasifikasikan data. Akurasi dapat ditentukan dengan membagi jumlah total sampel dengan jumlah prediksi yang benar. Di sisi lain, recall mengukur berapa banyak prediksi yang menguntungkan berdasarkan pengamatan kelas yang benar [20].

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

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

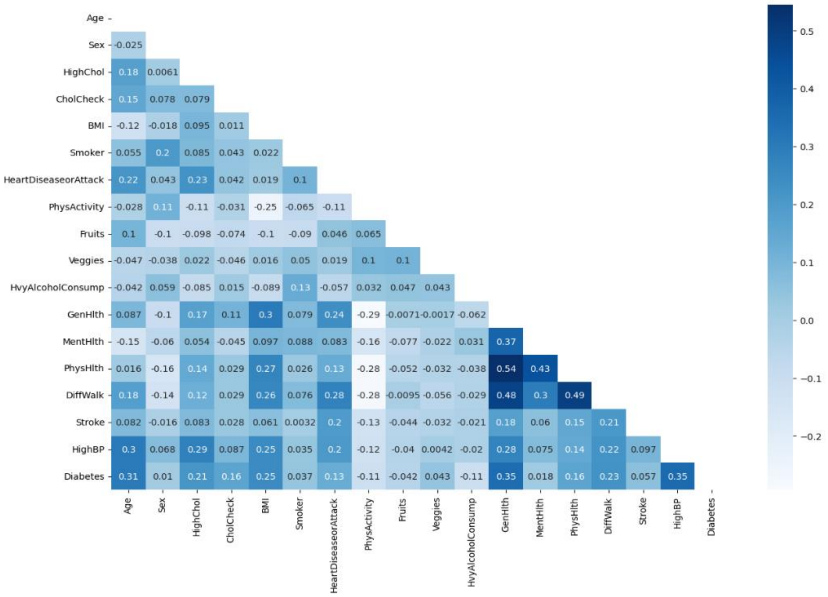
3. HASIL DAN PEMBAHASAN

3.1 Data Hasil Processing

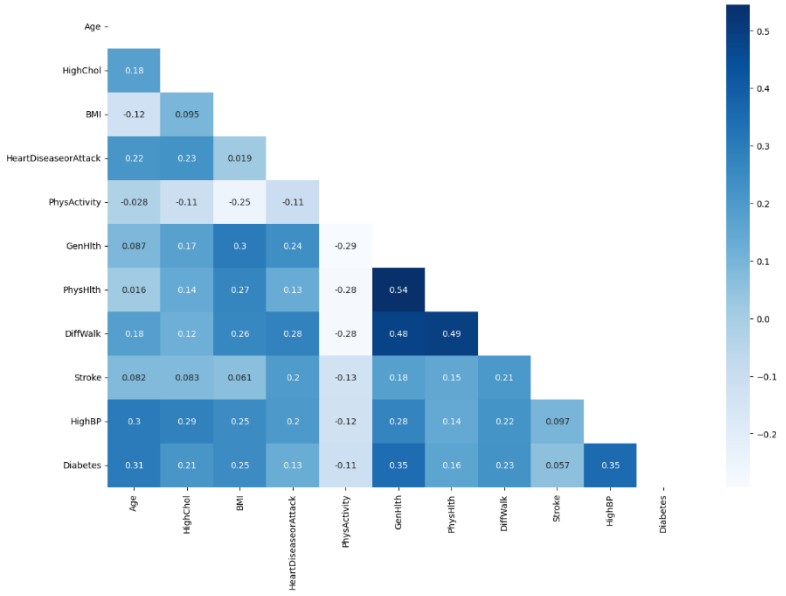
Aspek penting dari persiapan data adalah prosedur ini. Tujuan dari penghapusan fitur adalah untuk membuat kumpulan data lebih sesuai untuk analisis atau pemodelan dengan menghapus fitur yang tidak diperlukan. Dengan berapa

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

proporsi data yang digunakan untuk pelatihan dan pengujian.70% untuk data pelatihan dan 30% untuk data pengujian.Overfitting dan kompleksitas yang berlebihan adalah dua masalah yang dapat dicegah dengan prosedur ini.



Gambar 4. Data mentah



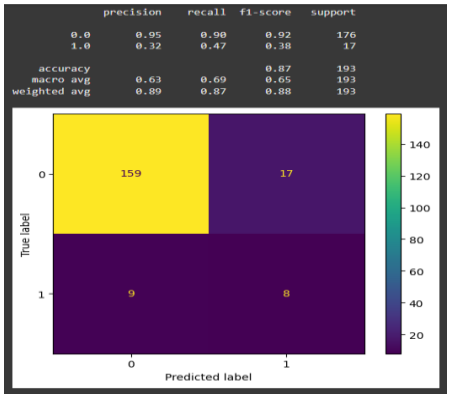
Gambar 5. Data setelah handling missing value

Tujuan pemisahan dataset, seperti yang ditunjukkan pada Gambar 4 dan Gambar 5, adalah untuk memfasilitasi pelatihan dan pengujian model dengan memisahkan data menjadi karakteristik input dan tujuan output. Opsi stratify=y digunakan untuk melakukan ini, dan menjaga distribusi kelas target yang tidak seimbang agar proporsional baik dalam set data pelatihan maupun pengujian. Untuk mencegah model menjadi bias karena variasi ukuran, normalisasi data dilakukan. Misalnya, jika fitur BMI memiliki nilai yang sangat besar, fitur tersebut mungkin mengalahkan fitur lain dengan skala yang lebih kecil. Untuk meningkatkan kinerja model dan membuat pelatihan lebih andal, StandardScaler digunakan untuk menstandardisasi semua fitur agar memiliki skala yang sama. Langkah kedua melibatkan pembuatan struktur DataFrame bernama accuracy_model untuk menyimpan data kinerja model setelah pelatihan. Dua kolom DataFrame ini adalah "Model" untuk menyimpan nama model dan "Accuracy" untuk menyimpan akurasi evaluasi setiap model. Model akan dilatih menggunakan data pelatihan yang dinormalkan (X_train_scaled, y_train), dan kinerjanya akan dievaluasi menggunakan metrik akurasi menggunakan data uji yang dinormalkan (X_test_scaled, y_test). Pada tahap selanjutnya, DataFrame accuracy_model akan memfasilitasi penyajian tabel dari temuan akhir, yang memungkinkan perbandingan kinerja model yang efektif.

3.2 Hasil Naïve Bayes

Hasil *confussion matrix* pada algoritma *Naïve Bayes* terlihat pada Gambar 6, sebagai berikut.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

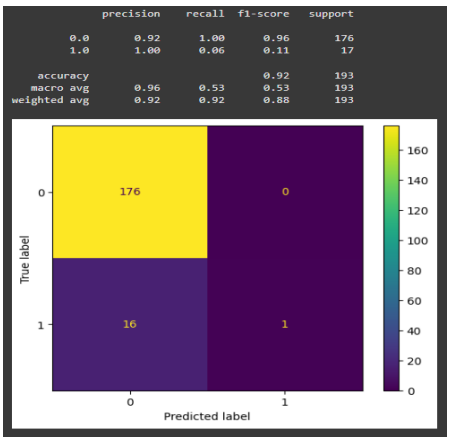


Gambar 6. Confussion matrix *Naïve Bayes*

Pada Gambar 6 representasi visual dari evaluasi Model Bernoulli Naive Bayes berkenaan dengan akurasi, presisi, recall, dan skor F1. Kelas 1 (goresan) memiliki presisi hanya 0,32 sedangkan kelas 0 (tanpa goresan) mencapai 0,95. Dengan kata lain, prediksi kelas 0 lebih akurat daripada prediksi kelas 1 menurut model tersebut. Model tersebut unggul dalam mengidentifikasi kelas 0 tetapi kesulitan dengan kelas 1, seperti yang terlihat dari recall sebesar 0,90 untuk kelas 0 dan 0,47 untuk kelas 1. Dengan skor F1 sebesar 0,92 untuk kelas 0 dan 0,38 untuk kelas 1, model tersebut jelas unggul dalam mengelola kelas mayoritas (kelas 0) karena keseimbangan yang harmonis antara recall dan akurasi

3.3 Random Forest

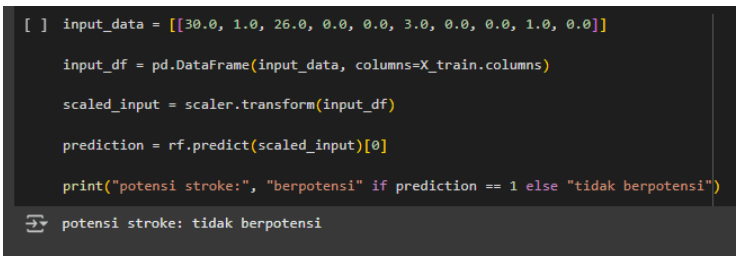
Hasil *Confussion matrix* pada algoritma random forest, terlihat pada Gambar 7.



Gambar 7. Confussion matrix *Random forest*

Pada Gambar 7 grafik menunjukkan bahwa kumpulan data baru (X_train_rf) digunakan untuk melatih Model Hutan Acak, dan kumpulan data uji (X_test_rf) digunakan untuk penilaian. Performa model ditunjukkan dalam hasil evaluasi menggunakan beberapa ukuran. Kelas 0 (tanpa stroke) memiliki presisi 0,92 dan kelas 1 (stroke) memiliki presisi 1,00, yang menunjukkan bahwa model secara akurat memprediksi kelas 0, mayoritas, tanpa menghasilkan positif palsu untuk kelas 1. Pada akhirnya, model mendapat skor 0,92. Meskipun akurasi ini mengesankan, akurasi ini gagal menangkap performa model pada kelas minoritas, faktor penting dalam diagnosis stroke. Total satu positif benar dan seratus tujuh puluh enam negatif benar dihasilkan oleh matriks kebingungan, dengan kelompok pertama terdiri dari data kelas 0 dan yang terakhir terdiri dari data kelas 1. Tidak ada prediksi yang salah dari kelas 0 dan 1, namun ada enam belas prediksi yang salah dari kelas 1 dan 0.

3.4 Implementasi



Gambar 8. Pengimplementasian data pada kedua algoritma

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

Prosedur untuk membuat prediksi risiko stroke menggunakan model Random Forest yang telah dilatih sebelumnya ditunjukkan pada gambar. Usia, jenis kelamin, indeks massa tubuh (IMT), dan variabel karakteristik lainnya membentuk array yang berfungsi sebagai data input. Langkah berikutnya adalah mengubah data input menjadi DataFrame dengan menambahkan kolom yang sesuai dengan data pelatihan. Setelah menyesuaikan StandardScaler dengan data pelatihan, data input harus dinormalisasi. Langkah berikutnya adalah menerapkan model Random Forest ke data input yang dinormalisasi untuk membuat prediksi. Hasil ramalan dapat dinyatakan sebagai 0 (tidak ada kemungkinan stroke) atau 1 (kemungkinan stroke). Di sini, temuan ramalan menunjukkan bahwa orang dengan data ini tidak mungkin terkena stroke. Untuk menampilkan hasil prediksi dalam bahasa yang mudah dipahami, kalimat ini ditampilkan dalam output menggunakan logika dasar yang memanfaatkan if-else. Semua langkah yang terlibat dalam membuat ramalan menggunakan model pembelajaran mesin mulai dari membersihkan dan mengatur data hingga menarik kesimpulan ditunjukkan di sini.

4 KESIMPULAN

Dari persiapan awal hingga penilaian akhir, penelitian ini berhasil dengan sangat baik. Penelitian menemukan bahwa algoritma random forest memiliki tingkat keberhasilan lebih baik dengan score precision 92%, Recall 92% dan f1-score 88% dalam mengklasifikasi penyakit stroke, sedangkan metode naïve bayes mencapai tingkat keberhasilan score precision 89%, Recall 87% dan f1-score1 88% . Temuan ini seharusnya memotivasi penyelidikan lebih lanjut tentang kemungkinan membandingkan Random Forest dan algoritma naïve bayes dengan algoritma pembelajaran mesin lainnya di masa mendatang. Dengan melakukan hal ini, kita mungkin memiliki pemahaman yang lebih komprehensif tentang kinerja algoritma dan memperoleh lebih banyak wawasan tentang model mana yang akan digunakan dalam mengklasifikasi penyakit stroke untuk deteksi dini pada dunia medis.

REFERENCES

[1] G. Ciaburro and G. Iannace, “Machine learning-based algorithms to knowledge extraction from time series data: A review,” *Data (Basel)*, vol. 6, no. 6, pp. 1–30, 2021, doi: 10.3390/data6060055.
[2] Wijoyo A, Saputra A, Ristanti S, Sya’ban S, Amalia M, and Febriansyah R, “Pembelajaran Machine Learning,” *OKTAL (Jurnal Ilmu Komputer dan Science)*, vol. 3, no. 2, pp. 375–380, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal/article/view/2305
[3] A. Jariah and I. Suaidah, “Efektivitas Penerapan Model Pembelajaran Inquiry Terbimbing Berbantuan Pictorial Riddle terhadap Pemahaman Konsep Fisika Peserta Didik di SMAN 14 Bulukumba,” *Jurnal Pendidikan Tambusai*, vol. 9, no. 2, pp. 26503–26510, 2025.
[4] A. Q. Khan, M. Matskin, R. Prodan, C. Bussler, D. Roman, and A. Soylu, “Cloud storage tier optimization through storage object classification,” *Computing*, vol. 106, no. 11, pp. 3389–3418, 2024, doi: 10.1007/s00607-024-01281-2.
[5] L. Dewi and E. Fitraneti, “Stroke Iskemik,” *Scientific Journal*, vol. 3, no. 6, pp. 365–374, 2024, doi: 10.56260/sciena.v3i6.173.
[6] T. Inui *et al.*, “The role of micronutrients in ageing asia: What can be implemented with the existing insights,” *Nutrients*, vol. 13, no. 7, pp. 1–27, 2021, doi: 10.3390/nu13072222.
[7] J. Bekker and J. Davis, *Learning from positive and unlabeled data: a survey*, vol. 109, no. 4. Springer US, 2020. doi: 10.1007/s10994-020-05877-5.
[8] D. J. Hand, P. Christen, and N. Kirielle, “F*: an interpretable transformation of the F-measure,” *Mach Learn*, vol. 110, no. 3, pp. 451–456, 2021, doi: 10.1007/s10994-021-05964-1.
[9] A. D. P. Harefa, K. I. Nahampun, P. Ritonga, and Ivo. Andre, “Penerapan Metode Naive Bayes Untuk Memprediksi Hasil Belajar Peserta Didik ‘Studi Kasus SD Swasta Methodist-1 Rantauprapat,’” *Journal Computer Science and Information Technology(JCoInT)*, vol. 5, no. 2, pp. 238–248, 2024.
[10] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, “Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur Xgboost Untuk Klasifikasi Kanker Payudara,” *TIN: Terapan Informatika Nusantara*, vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.
[11] K. Gajowniczek and M. Dudziński, “Influence of Explanatory Variable Distributions on the Behavior of the Impurity Measures Used in Classification Tree Learning,” *Entropy*, vol. 26, no. 12, pp. 1–35, 2024, doi: 10.3390/e26121020.
[12] A. Durap, “Data-driven models for significant wave height forecasting: Comparative analysis of machine learning techniques,” *Results in Engineering*, vol. 24, no. November, p. 103573, 2024, doi: 10.1016/j.rineng.2024.103573.
[13] P. Anggraini and W. Winarsih, “Komparasi Naïve Bayes, Support Vector Machine, Dan Random Forest Dalam Analisis Sentimen Aplikasi Shopee Di Google Play Store,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 3, pp. 4451–4457, 2025, doi: 10.36040/jati.v9i3.13675.
[14] N. Ayub, Tayyaba, S. Hussain, S. S. Ullah, and J. Iqbal, “An Efficient Optimized DenseNet Model for Aspect-Based Multi-Label Classification,” *Algorithms*, vol. 16, no. 12, pp. 1–30, 2023, doi: 10.3390/a16120548.
[15] A. Widianti and I. Pratama, “Penanganan Missing Values dan Prediksi Data Timbunan,” *RABIT : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 9, no. 2, pp. 242–251, 2024.
[16] K. A. Barchard and J. A. Russell, “Distorted correlations among censored data: causes, effects, and correction,” *Behav Res Methods*, vol. 56, no. 3, pp. 1207–1228, 2024, doi: 10.3758/s13428-023-02086-5.
[17] C. A. Ramezan, T. A. Warner, A. E. Maxwell, and B. S. Price, “Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data,” *Remote Sens (Basel)*, vol. 13, no. 3, pp. 1–27, 2021, doi: 10.3390/rs13030368.
[18] H. Oktavianto, H. W. Sulistyo, G. Wijaya, D. Irawan, and G. Abdurrahman, “Analisis Perbandingan Decision Tree dan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2807−2814
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6771

Random Forest Pada Klasifikasi Teks Data Kesehatan," *Bina Insani Ict Journal*, vol. 11, no. 1, p. 56, 2024, doi: 10.51211/biict.v11i1.2928.

[19] J. Kozak, B. Probierz, K. Kania, and P. Juszczuk, "Preference-Driven Classification Measure," *Entropy*, vol. 24, no. 4, pp. 1–24, 2022, doi: 10.3390/e24040531.

[20] K. W. C. Wahyuditomo Imam; Sutrisno, Sutrisno, "Implementasi Integrasi K-Means dan Naïve Bayes dalam Identifikasi Tingkat Risiko Reksa Dana," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 7, pp. 3094–3102, 2021, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/9487/4284