Komparasi Teknik Bagging Dan Adaboost Pada Decision Tree Dan Naive Bayes Untuk Prediksi Stroke

Hafsah Mukaromah*[1], Wasilah[2]
[1,2]Program Studi Magister Teknik Informatika, Institut Informatika dan Bisnis Darmajaya; Jl. ZA. Pagar Alam No.93, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, 35141, Indonesia, telp (0721) 787214
e-mail: ***[1]hafsah.2221210013@mail.darmajaya.ac.id,** [2]wasilah@darmajaya.ac.id

Abstrak

Stroke, yang juga dikenal sebagai cerebrovascular accident (CVA), adalah kondisi di mana terjadi gangguan mendadak dalam fungsi otak akibat masalah peredaran darah, yang dapat mengakibatkan kelumpuhan atau bahkan kematian sel otak. Stroke terdiri dari dua jenis yaitu iskemik, yang disebabkan oleh penyumbatan pembuluh darah, dan hemoragik, yang disebabkan oleh pendarahan ke otak. Di Indonesia, stroke menjadi penyebab kematian utama dengan tingkat kejadian yang terus meningkat. Oleh karena itu, upaya pencegahan dan pengobatan dini sangat penting dalam penanganan kondisi ini. Data mining dan machine learning telah menjadi alat yang penting dalam memprediksi risiko stroke. Dalam penelitian ini, teknik ensemble, khususnya bagging dan adaboost, diterapkan pada algoritma decision tree dan naive bayes untuk meningkatkan akurasi dalam memprediksi stroke. Hasil penelitian menunjukkan bahwa penggunaan teknik ensemble, terutama adaboost, mampu secara signifikan meningkatkan kinerja algoritma naive bayes, dengan peningkatan akurasi hingga 7,42%. Kombinasi algoritma decision tree dengan bagging memberikan akurasi tertinggi dalam memprediksi stroke, mencapai 96,91%, diikuti oleh kombinasi decision tree dengan adaboost dan naive bayes dengan adaboost. Temuan ini menunjukkan bahwa penggunaan teknik ensemble dapat secara signifikan meningkatkan kinerja algoritma dalam memprediksi penyakit stroke, dengan fokus pada penggunaan Adaboost untuk algoritma Naive Bayes dan metode bagging untuk decision tree.
***Kata kunci**—stroke, decision tree, naïve bayes, adaboost, bagging*

Abstract

Stroke, also known as a cerebrovascular accident (CVA), is a condition in which there is a sudden disruption in brain function due to circulatory problems, which can result in paralysis or even death of brain cells. Stroke consists of two types: ischemic, caused by blood vessel blockage, and hemorrhagic, caused by bleeding into the brain. In Indonesia, stroke is a leading cause of death with incidence rates continuing to rise. Therefore, prevention efforts and early treatment are crucial in managing this condition. Data mining and machine learning have become important tools in predicting stroke risk. In this study, ensemble techniques, particularly bagging and Adaboost, were applied to decision tree and naive Bayes algorithms to improve accuracy in predicting stroke. The research results indicate that the use of ensemble techniques, especially Adaboost, significantly enhances the performance of the naive Bayes algorithm, with accuracy increasing by 7.42%. The combination of decision tree algorithm with bagging provides the highest accuracy in predicting stroke, reaching 96.91%, followed by the combination of decision tree with Adaboost and naive Bayes with Adaboost. These findings demonstrate that the use of ensemble techniques can significantly improve the

performance of algorithms in predicting stroke, with a focus on utilizing Adaboost for the naive Bayes algorithm and bagging method for decision trees.
Keywords— *stroke, decision tree, naïve bayes, adaboost, bagging*

1. PENDAHULUAN

Stroke atau cerebrovascular accident (CVA) adalah keadaan dimana fungsi otak hilang akibat gangguan mendadak pada suplai darah ke otak. Gangguan ini terjadi karena masalah sirkulasi darah di otak yang menyebabkan kelumpuhan atau kematian.[1] Sel otak mengalami kematian akibat penyumbatan saluran yang menyediakan nutrisi dan oksigen ke otak. [2] Stroke dibagi menjadi 2 jenis, yaitu stroke iskemik, yang biasanya disebabkan oleh penyumbatan pembuluh darah, sementara stroke hemoragik adalah jenis stroke yang disebabkan oleh pendarahan langsung ke otak atau ke dalam ruang antara membran otak. Informasi dari Organisasi Kesehatan Dunia (WHO) menunjukkan bahwa ada 10 penyebab kematian utama di Indonesia. Stroke menempati urutan pertama, dengan 131,8 kematian per 100.000 orang. Kedua adalah penyakit jantung koroner iskemik atau penyebab gagal jantung, dengan 95,68 kasus. Urutan ketiga, dengan selisih yang signifikan, adalah diabetes melitus dengan 40,78 kasus. Keempat, tuberkulosis (TB) meningkat sebanyak 33,24 kasus. Dengan 33,06 kasus, sirosis hati berada tepat di bawah persentil kelima. Urutan terakhir adalah kematian bayi baru lahir dengan 16,77 kasus.[3] Angka kejadian stroke di Indonesia menurut data Riskedas pada tahun 2013 adalah 7 per 1.000 pasien stroke. Angka kejadian ini meningkat pada tahun 2018 menjadi 10,9 per 1.000 pasien stroke. Banyak faktor yang masuk dalam empat layanan prioritas berdasarkan data Kementerian Kesehatan tahun 2022, termasuk fakta bahwa tiga dari setiap 1.000 orang berisiko terkena stroke setiap tahunnya. 15% korban stroke terancam mati dan 65% korban stroke terancam mengalami kecacatan.[4]

Permasalahan stroke di Indonesia membutuhkan perhatian serius mengingat jumlah kasus yang terus meningkat dan tingginya angka kematian.[5] Diperlukan upaya untuk melakukan prediksi stroke secara dini guna mendukung tindakan pencegahan dan pengobatan awal. Salah satu cara untuk mengatasi masalah ini adalah dengan mengumpulkan informasi dari data stroke, yang dapat digunakan untuk mengembangkan model baru yang mendetail tentang profil pasien dan menentukan risiko stroke mereka. Data mining adalah suatu metode yang terlibat dalam ekstraksi informasi yang berharga dari kumpulan data besar. Hasil dari proses data mining ini dapat dimanfaatkan untuk meningkatkan proses pengambilan keputusan di masa mendatang.[6] Salah satu pendekatan dalam pengenalan pola dalam data mining adalah melalui penerapan machine learning, di mana komputer mempelajari pola dari data tertentu dan membentuk model. Dalam machine learning, terdapat dua pendekatan utama: supervised learning dan unsupervised learning. Supervised learning, termasuk klasifikasi, digunakan ketika hasil yang diinginkan atau yang diperkirakan sudah diketahui. Klasifikasi merupakan salah satu teknik dalam data mining yang digunakan untuk mengeksplorasi informasi penting dari suatu set data.

Teknik klasifikasi mampu secara otomatis memprediksi kelas dari data yang belum terklasifikasi.[7] Decision Tree dan Naïve Bayes adalah dua metode klasifikasi yang umum digunakan dalam data mining untuk memprediksi penyakit. Decision Tree menghasilkan keputusan dalam bentuk struktur pohon yang fleksibel, yang mampu meningkatkan kualitas keputusan. Keuntungan utama dari Decision Tree adalah fleksibilitasnya yang memungkinkan untuk peningkatan kualitas keputusan yang

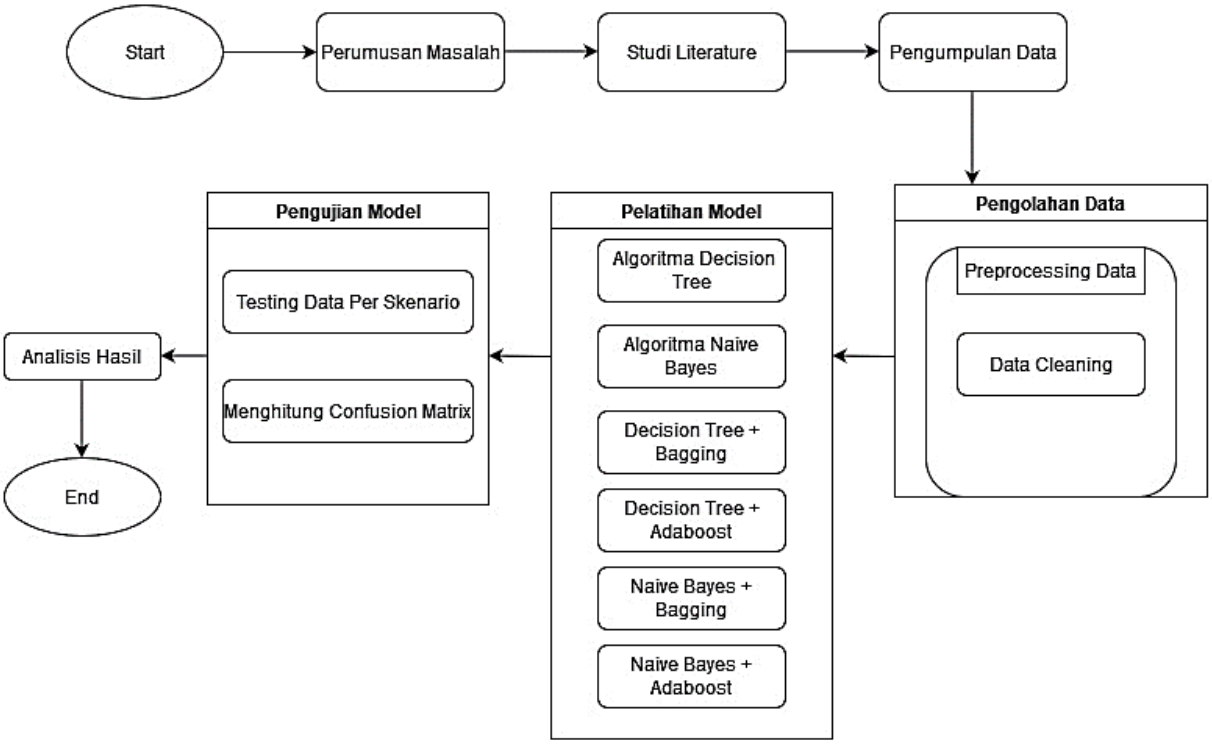
dihasilkan.[8] Sementara itu, Naïve Bayes merupakan metode klasifikasi probabilitas sederhana yang mengestimasi probabilitas berdasarkan frekuensi dan kombinasi nilai dari dataset. Namun, Naïve Bayes hanya mendukung tipe data atribut diskrit dan memiliki keterbatasan pada atribut kontinu. Selain itu, atribut tersebut dapat memberikan kontribusi kepada atribut yang diprediksi.[9] Metode naïve bayes juga membutuhkan jumlah data pelatihan yang relatif kecil untuk menentukan parameter yang diperlukan dalam proses klasifikasi.[10] Teknik ensemble, seperti bagging dan boosting, telah diterapkan untuk meningkatkan akurasi algoritma decision tree dan naïve bayes dalam memprediksi stroke. Teknik Bagging merupakan teknik yang sukses untuk menangani dataset yang tidak seimbang.[11] Bagging mendukung algoritma klasifikasi yang tidak stabil seperti decision trees, sementara boosting bertujuan untuk meningkatkan kinerja klasifikasi dengan memanfaatkan kekuatan kolektif dari model lemah. dalam beberapa situasi, AdaBoost cenderung lebih tahan terhadap masalah overfitting jika dibandingkan dengan algoritma pembelajaran lainnya.[12] Beberapa penelitian yang telah dilakukan sebelumnya terkait prediksi penyakit stroke menggunakan algoritma decision tree, naïve bayes dan teknik ensemble berjudul "Komparasi Penerapan Metode Bagging dan Adaboost pada Algoritma C4.5 untuk Prediksi Penyakit Stroke" di teliti oleh [13] menghasilkan nilai akurasi algoritma C4.5 sebesar 92,87%. Namun, setelah menerapkan teknik bagging, akurasi meningkat menjadi 95,02%, sedangkan setelah menerapkan metode Adaboost, akurasi mencapai 94,63%. Perbandingan antara teknik Bagging dan Adaboost pada algoritma C4.5 membuktikan peningkatan dan perbaikan kinerja klasifikasi. Akurasi algoritma C4.5 meningkat sebesar 3% dan 2% setelah digunakan metode bagging dan Adaboost secara berturut-turut. Selanjutnya penelitian yang dilakukan oleh [14] berjudul *"Analyzing the Performance of Stroke Prediction using ML Classification Algorithms"* dengan dataset berjumlah 5110 data. Hasil uji coba pada 6 algoritma klasifikasi yang dipilih, algoritma Naïve Bayes Classification memiliki kinerja terbaik dengan akurasi 82%. Sedangkan algoritma Logistic Regression memiliki akurasi 78%, Decision Tree Classification 66%, Random Forest Classification 73%, K-Nearest Neighbors Classification 80%, dan Support Vector Machine sebesar 80%. Kemudian penelitian berikutnya yang berjudul [15] hasil penelitian yang telah dilakukan yaitu dengan membagi dataset menjadi 60% data training dan 40% data testing maka dapat disimpulkan bahwa algoritma C4.5 memiliki performa yang lebih baik yaitu dengan tingkat akurasi sebesar 95%.sedangkan algoritma naïvebayes mendapatkan tingkat akurasi sebesar 91%. Selanjutnya penelitian dengan judul [1] hasil pengujian 10 fold cross validation, diperoleh bahwa algoritma decision tree mencapai akurasi tertinggi sebesar 0.953. Sedangkan Adaboost mencapai 0.915% dan Random Forest mencapai 0.950%. Dari hasil penelitian ini, dapat disimpulkan bahwa pohon keputusan efektif untuk melakukan klasifikasi stroke. Namun, penting untuk dicatat bahwa hasil ini bergantung pada dataset yang digunakan. Studi perbandingan tentang penerapan teknik bagging dan adaboost pada algoritma decision tree dan naïve bayes untuk prediksi stroke telah menunjukkan peningkatan signifikan dalam kinerja klasifikasi.

Penelitian tentang penyakit stroke dari sudut pandang bidang komputer melibatkan penerapan teknik dan algoritma komputasi untuk menganalisis data medis, mendiagnosis, memprediksi risiko, dan bahkan merancang intervensi yang lebih efektif. Oleh karena itu, penelitian ini sangat penting untuk meningkatkan pemahaman tentang penyakit stroke, meningkatkan diagnosis dan pengobatan, serta mengurangi angka kejadian dan dampaknya pada individu dan masyarakat. Maka dari itu, Penelitian ini bertujuan untuk membandingkan algoritma klasifikasi decision tree dan naïve bayes serta penerapan teknik ensemble untuk menentukan kombinasi mana yang mencapai akurasi tertinggi dalam prediksi stroke, dengan harapan meningkatkan akurasi prediksi dan mengevaluasi kinerja algoritma decision tree dan naïve bayes dalam memprediksi stroke.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini merupakan jenis penelitian eksperimen. Dimana data yang dikumpulkan akan dilakukan pengujian menggunakan model algoritma klasifikasi serta penerapan teknik ensemble. Proses penelitian ini dimulai dari merumuskan masalah, studi literature, pengumpulan data, pengolahan data, pelatihan model yang terdiri dari enam skenario pengujian, pengujian model yang terdiri dari testing data per-skenario dan menghitung confusion matrix dan terakhir yaitu analisis hasil. Ilustrasi pada Gambar 1 menggambarkan alur keseluruhan dalam penelitian ini.



Gambar 1. Alur Penelitian

2.1.1 Perumusan Masalah

Tahap ini merupakan langkah awal sebelum memulai penelitian, yaitu tahap identifikasi permasalahan. Pada tahap ini, pemahaman terhadap permasalahan penelitian dikemukakan berdasarkan konteks penelitian mengenai prediksi penyakit stroke. Penelitian ini menggunakan algoritma Decision Tree dan Naïve Bayes serta menerapkan teknik ensemble seperti Bagging dan Adaboost untuk melakukan prediksi penyakit stroke. Dengan menggunakan metode ini, dapat dipastikan apakah seorang pasien memiliki risiko mengalami stroke atau tidak dengan tingkat akurasi yang tinggi.

2.1.2 Studi Literature

Langkah kedua melibatkan pencarian dan pengumpulan makalah dari berbagai sumber yang berkaitan dengan penelitian, yang akan digunakan sebagai referensi. Sumber referensi ini bisa berupa buku, jurnal, dan e-book yang relevan dengan topik penelitian atau metode yang digunakan dalam penelitian. Proses studi literatur ini bertujuan untuk memperoleh pemahaman yang mendalam mengenai penelitian terdahulu yang berkaitan dengan stroke serta metode penelitian yang relevan.

2.1.3 Pengumpulan Data

Dataset Stroke Prediction adalah sebuah kumpulan data yang diperoleh dari Kaggle, terdiri dari 5110 data mentah. Dataset ini terdiri dari 11 atribut prediktor dan satu atribut target yang merujuk pada kejadian stroke. Atribut stroke mencakup informasi tentang pasien yang mengalami stroke dan yang tidak. Proporsi pasien yang tidak mengalami stroke mencapai 95%, sementara pasien yang mengalami stroke hanya 5%. Dengan demikian, berdasarkan penjelasan tersebut, dataset ini dapat diklasifikasikan sebagai dataset yang tidak seimbang karena jumlah pasien yang tidak mengalami stroke lebih dominan dibandingkan dengan yang mengalami stroke.

2.1.4 Pengolahan Data

Tahap pengolahan data merupakan bagian penting dari proses analisis data yang melibatkan serangkaian langkah untuk membersihkan, mengatur, dan menyiapkan data mentah sehingga siap digunakan untuk analisis lebih lanjut. Pada tahap ini, dilakukan proses preprocessing data untuk mengubah data mentah menjadi format yang dapat dengan mudah diolah menggunakan teknik data mining. Beberapa atribut masih memiliki data yang tidak konsisten, misalnya "N/A". Oleh karena itu, akan dilakukan pembersihan data atau data cleaning untuk menghapus data yang tidak sesuai dan menggantinya dengan nilai rata-rata dari atribut yang memiliki nilai "N/A" tersebut. Dalam dataset prediksi stroke, terdapat 201 data yang memiliki nilai "N/A" pada atribut BMI.

2.1.5 Pelatihan Model

Setelah melewati tahap preprocessing atau pengolahan data, langkah selanjutnya adalah tahap pelatihan model. Pada tahap keempat ini, klasifikasi akan dilakukan menggunakan algoritma Decision Tree dan Naïve Bayes, dengan menerapkan teknik bagging dan adaboost. Proses ini melibatkan pembagian klasifikasi ke dalam enam skenario yang berbeda.

Tabel 1. Skenario Pengujian

Decision Tree	Naïve Bayes	Decision Tree + Bagging	Decision Tree + Adaboost	Naïve Bayes + Bagging	Naïve Bayes + Adaboost
Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decision tree saja tanpa menerapkan teknik *Bagging* atau *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes saja tanpa menerapkan teknik *Bagging* atau *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decision tree dan teknik bagging	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decision tree dan teknik *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes dan teknik bagging	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes dan *Adaboost*

2.1.6 Pengujian Model

Tahap terakhir dari penelitian ini merupakan fase akhir, yang mencakup pengujian model pada enam skenario klasifikasi menggunakan data yang telah di-label. Sebelum melakukan pengujian model, perlu dilakukan proses pengolahan data terlebih dahulu untuk memastikan hasil pengujian optimal. Proses pengolahan data ini dilakukan melalui tahap preprocessing. Setelah itu, data akan dimasukkan ke dalam model atau algoritma klasifikasi untuk mengevaluasi kinerja model tersebut. Pengujian model dilakukan menggunakan metode confusion matrix. Sebelum melakukan pengujian, data dipisahkan menjadi data latih dan data uji. Pengujian dilakukan dengan menggunakan k-fold cross-validation, di mana 10 sampel data diambil secara acak untuk

mengevaluasi model. Dari berbagai sampel data tersebut, model yang memberikan hasil optimal dipilih dengan mencari rata-rata dari 10 sampel tersebut untuk mendapatkan nilai akurasi, presisi, spesifisitas, dan sensitivitas.

2.1.7 Analisis Hasil

Tahap analisis hasil difokuskan pada membandingkan hasil dari masing-masing skenario yang telah melewati tahap pengujian sebelumnya. Enam skenario ini melibatkan pengujian berbagai kombinasi algoritma, termasuk pengujian dengan algoritma decision tree, algoritma naïve bayes, algoritma decision tree dengan teknik bagging, algoritma decision tree dengan teknik adaboost, algoritma naïve bayes dengan teknik bagging, dan algoritma naïve bayes dengan teknik adaboost. Dari pengujian pada keenam skenario tersebut, dihasilkan nilai presisi, spesifisitas, sensitivitas, dan akurasi. Dengan membandingkan nilai-nilai tersebut, kita dapat menentukan metode mana yang memiliki dampak signifikan dalam meningkatkan performa klasifikasi. Selain itu, tahap analisis juga mencakup membandingkan hasil penelitian ini dengan penelitian lain yang menggunakan metode serupa.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

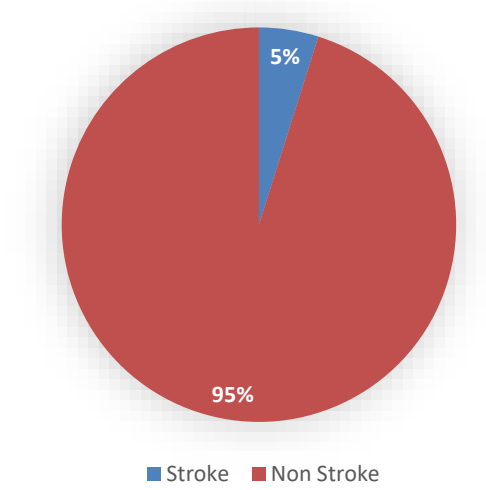
Pada penelitian ini dataset yang digunakan yaitu dataset stroke prediction yang diunduh dari situs kaggle.com. Jumlah data awal sebelum proses preprocessing adalah sebanyak 5110 data. Dataset ini terdiri dari 12 atribut, dimana 11 di antaranya berfungsi sebagai variabel prediktor, meliputi id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi (body mass index), dan smoking_status. Sementara itu, satu atribut lainnya berperan sebagai variabel target atau label, yaitu stroke. Variabel target ini memiliki dua nilai output, yaitu 0 dan 1. Nilai 0 menunjukkan bahwa pasien tidak terkena stroke, sementara nilai 1 menandakan bahwa pasien terkena stroke. Berikut gambar 2 menunjukan dataset awal atau mentah yang belum dilakukan proses preprocessing data.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	22869	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	20221	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	10592	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	17123	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	17412	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	18621	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	7009	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	9439	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	7615	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	5857	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	8043	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	12046	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	10451	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	21984	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	21409	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	16741	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	19161	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	22129	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	8922	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	21708	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	19394	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	23329	48.9	never smoked	1
68794	Female	79	0	0	Yes	Self-employed	Urban	2287	26.6	never smoked	1

Gambar 2. Dataset yang digunakan

Berdasarkan gambar 2 diatas, atribut stroke menunjukkan bahwa terdapat 4861 data pasien yang tidak mengalami stroke, sedangkan jumlah data pasien yang mengalami stroke hanya 249. Dengan demikian, dapat disimpulkan bahwa jumlah pasien yang tidak mengalami stroke jauh lebih banyak dibandingkan dengan jumlah pasien yang mengalami stroke. Hal ini menandakan ketidakseimbangan kelas data pada atribut stroke. Gambar 3 di bawah ini memberikan ilustrasi perbandingan jumlah data pasien yang mengalami stroke dan yang tidak. Persentase pasien yang

tidak mengalami stroke mencapai 95%, sementara persentase pasien yang mengalami stroke hanya sebesar 5%.



Gambar 3. Persentase Pasien Stroke

3.2 Pengolahan Data

Pada dataset stroke prediction yang berjumlah 5110 data didalamnya terdapat beberapa data noise. Noise adalah data yang berisi nilai-nilai yang salah atau anomali, yang biasanya disebut juga outlier. [16] Data noise merupakan data yang mengalami kerusakan, seperti data duplikat, data tidak konsisten, data yang hilang/missing value, data tidak beraturan dan data outlier yang dapat mempengaruhi nilai akurasi pada proses klasifikasi.

3.2.1 Data Cleaning

Data Cleaning merupakan serangkaian tindakan yang melibatkan identifikasi, penanganan, dan perbaikan masalah dalam dataset. Tujuannya adalah menghasilkan dataset yang lebih bersih, konsisten, dan sesuai untuk analisis atau pemodelan.[17] Proses pembersihan data dilakukan pada dataset stroke prediction yang memiliki nilai tidak konsisten N/A dan missing value. Pada dataset ini ada atribut yang memiliki nilai N/A sebanyak 201 data yaitu atribut bmi. Data tersebut merupakan data noise, sehingga sebelum masuk ke tahap pelatihan model harus dihapus atau diganti agar tidak mempengaruhi nilai akurasi. Data yang tidak konsisten berupa nilai N/A akan diubah menjadi nilai rata-rata dari atribut BMI. Untuk menghandling atribut bmi yang memiliki nilai N/A, ada 2 tools yang digunakan pada penelitian ini yaitu Microsoft excel dan aplikasi rapid miner.

Setelah tahapan preprocessing data selesai dan data siap diolah, selanjutnya masuk ke tahap pelatihan model untuk mengetahui tingkat akurasi algoritma yang digunakan. Pada penelitian ini algoritma inti yang digunakan yaitu decision tree dan naïve bayes. Sedangkan teknik ensemble learning yang digunakan yaitu adaboost dan bagging. Pada pengujian 2 algoritma inti yang dikombinasi dengan teknik ensemble tersebut maka dibuat 6 skenario pengujian, diantaranya yaitu pengujian algoritma decision tree, naïve bayes, decision tree + bagging, decision tree + adaboost, naïve bayes + bagging dan naïve bayes + adaboost. Sehingga akan didapatkan 6 hasil akurasi.

3.3 Pelatihan Model

3.3.1 Klasifikasi Algoritma Decision Tree

Setelah menjalani tahap pengolahan data atau preprocessing pada *dataset stroke prediction*, langkah berikutnya adalah memasuki tahap pelatihan model. Berdasarkan skenario klasifikasi yang tercantum dalam tabel 1, model pertama yang dilatih menggunakan algoritma decision tree. Proses klasifikasi dilakukan menggunakan alat bantu rapid miner. Proses klasifikasi

terdiri dari dua tahap, yaitu tahap pelatihan (training) dan pengujian (testing). Dari proses klasifikasi ini, dihasilkan berbagai output, salah satunya adalah nilai akurasi dari algoritma decision tree.

3.3.2 Klasifikasi Algoritma Naïve Bayes

Pelatihan model kedua yaitu menggunakan algoritma naïve bayes. Algoritma ini adalah metode klasifikasi berbasis probabilitas yang menggunakan teorema Bayes. Langkah-langkah yang dilakukan diantaranya adalah pemahaman data, pembagian data, ekstraksi fitur, perhitungan probabilitas prior, perhitungan probabilitas likelihood, perhitungan probabilitas posterior, pemilihan kelas, evaluasi model dan implementasi. Proses klasifikasi menggunakan aplikasi rapid miner. Dari proses klasifikasi tersebut dihasilkan output salah satunya berupa nilai akurasi dari algoritma naïve bayes.

3.3.3 Klasifikasi Algoritma Decision Tree + Bagging

Pelatihan model ketiga yaitu penerapan metode bagging pada algoritma decision tree. Bagging merupakan salah satu teknik ensemble learning, bertujuan untuk meningkatkan performa algoritma decision tree. Sehingga menghasilkan nilai akurasi yang lebih tinggi setelah dioptimalkan menggunakan metode tersebut.

3.3.4 Klasifikasi Algoritma Decision Tree + Adaboost

Pelatihan model keempat yaitu menerapkan metode adaboost pada algoritma decision tree. Adaboost merupakan salah satu teknik ensemble yang berfungsi sebagai boosting pada algoritma decision tree. Penerapan adaboost pada decision tree bertujuan untuk meningkatkan kinerja algoritma decision tree sehingga menghasilkan nilai akurasi yang lebih tinggi.

3.3.5 Klasifikasi Algoritma Naïve Bayes + Bagging

Pelatihan model kelima yaitu menerapkan metode bagging pada algoritma naïve bayes. Pada proses ini algoritma naïve bayes dikombinasikan dengan metode bagging dengan tujuan untuk meningkatkan kinerja dari algoritma naïve bayes. Sehingga diperoleh nilai akurasi yang lebih tinggi setelah dioptimalkan dengan metode bagging.

3.3.6 Klasifikasi Algoritma Naïve Bayes + Adaboost

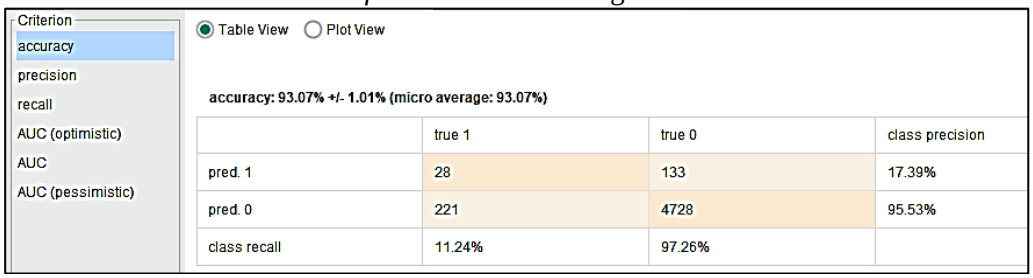
Pelatihan model yang keenam yaitu penerapan metode adaboost pada algoritma naïve bayes. Adaboost sebagai teknik ensemble berfungsi untuk meningkatkan kinerja dari algoritma naïve bayes.

3.4 Pengujian Model

Pengujian model pada *dataset stroke prediction* dilakukan untuk mengevaluasi kinerja model dalam memprediksi kemungkinan terjadinya stroke berdasarkan berbagai atribut yang ada dalam dataset.

3.4.1 Pengujian Algoritma Decision Tree

Tabel 2. *Confusion Matrix* dari Algoritma Decision Tree



	true 1	true 0	class precision
pred. 1	28	133	17.39%
pred. 0	221	4728	95.53%
class recall	11.24%	97.26%	

Berdasarkan tabel 2, dapat disimpulkan bahwa terdapat 28 pasien yang didiagnosis menderita stroke, sementara jumlah pasien yang sebenarnya tidak mengalami stroke atau non

stroke adalah 4728 orang. Terdapat juga 221 pasien stroke yang salah diidentifikasi sebagai non stroke, yang termasuk dalam kategori "Kesalahan Tipe II", sedangkan terdapat 133 pasien non stroke yang salah diidentifikasi sebagai stroke, yang termasuk dalam kategori "Kesalahan Tipe I". Kesalahan Tipe I mengacu pada situasi di mana pasien diprediksi menderita stroke, tetapi pada kenyataannya mereka tidak mengalami stroke. Sementara itu, Kesalahan Tipe II terjadi ketika pasien diprediksi tidak mengalami stroke, tetapi ternyata mereka sebenarnya mengalami stroke. Kesalahan Tipe II ini memiliki risiko yang tinggi karena dapat menyebabkan seseorang yang sebenarnya menderita stroke tidak mendapatkan diagnosis atau perawatan yang diperlukan.

3.4.2 Pengujian Algoritma Naïve Bayes

Tabel 3. *Confusion Matrix* dari Algoritma Naïve Bayes

PerformanceVector (Performance) | SimpleDistribution (Naive Bayes)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View / Plot View

accuracy: 87.14% +/- 1.14% (micro average: 87.14%)

	true 1	true 0	class precision
pred. 1	94	502	15.77%
pred. 0	155	4359	96.57%
class recall	37.75%	89.67%	

Dari data yang tercantum dalam tabel 3, dapat ditarik kesimpulan bahwa terdapat 94 pasien yang diprediksi menderita stroke, sementara pasien yang tidak mengalami stroke atau non stroke berjumlah 4359 orang. Selain itu, terdapat 155 pasien yang sebenarnya mengalami stroke tetapi salah diidentifikasi sebagai non stroke, yang termasuk dalam kategori "Type II Error". Di sisi lain, terdapat 502 pasien non stroke yang salah diidentifikasi sebagai stroke, yang masuk ke dalam kategori "Type I Error".

3.4.3 Pengujian Algoritma Decision Tree + Bagging

Tabel 4. *Confusion Matrix* dari Algoritma Decision Tree dan Bagging

PerformanceVector (Performance)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View / Plot View

accuracy: 96.91%

	true 1	true 0	class precision
pred. 1	92	1	98.92%
pred. 0	157	4860	96.87%
class recall	36.95%	99.98%	

Dari informasi yang tertera dalam tabel 4, dapat diketahui bahwa terdapat 92 pasien yang diidentifikasi menderita stroke, sementara jumlah pasien yang sebenarnya tidak mengalami stroke atau non stroke adalah 4860 orang. Selain itu, terdapat 157 pasien stroke yang salah diidentifikasi sebagai non stroke, yang masuk dalam kategori "Type II Error". Di sisi lain, hanya terdapat 1 pasien non stroke yang salah didiagnosa sebagai stroke, yang termasuk dalam kategori "Type I Error".

3.4.4 Pengujian Algoritma Decision Tree + Adaboost

Tabel 5. *Confusion Matrix* dari Algoritma Decision Tree dan Adaboost

PerformanceVector (Performance)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View / Plot View

accuracy: 96.58%

	true 1	true 0	class precision
pred. 1	102	28	78.46%
pred. 0	147	4833	97.05%
class recall	40.96%	99.42%	

Dari tabel 5 dapat diketahui bahwa terdapat 102 pasien yang terdiagnosis menderita stroke, 4833 pasien dinyatakan tidak menderita stroke, dan 147 pasien stroke yang salah diidentifikasi sebagai non-stroke, yang termasuk dalam kategori "Type II Error". Selain itu, 28 pasien non-stroke yang salah diprediksi sebagai stroke, yang masuk dalam kategori "Type I Error".

3.4.5 Pengujian Algoritma Naïve Bayes + Bagging

Tabel 6. *Confusion Matrix* dari Algoritma Naïve Bayes dan Bagging

PerformanceVector (Performance)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View / Plot View

accuracy: 87.22%

	true 1	true 0	class precision
pred. 1	95	499	15.99%
pred. 0	154	4362	96.59%
class recall	38.15%	89.73%	

Dari tabel 6 dapat disimpulkan bahwa terdapat 95 pasien yang terdiagnosis menderita stroke, 4362 pasien dinyatakan tidak menderita stroke, dan 154 pasien stroke yang salah diprediksi sebagai non-stroke, yang termasuk dalam kategori "Type II Error". Selain itu, 499 pasien non-stroke salah diprediksi sebagai stroke, yang masuk dalam kategori "Type I Error".

3.4.6 Pengujian Algoritma Naïve Bayes + Adaboost

Tabel 7. *Confusion Matrix* dari Algoritma Naïve Bayes dan Adaboost

PerformanceVector (Performance)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Table View / Plot View

accuracy: 94.56%

	true 1	true 0	class precision
pred. 1	5	34	12.82%
pred. 0	244	4827	95.19%
class recall	2.01%	99.30%	

Berdasarkan tabel 7 dapat ditarik kesimpulan bahwa terdapat 28 pasien yang teridentifikasi menderita stroke, 4827 pasien dinyatakan tidak menderita stroke, dan 244 pasien

stroke salah diprediksi sebagai non-stroke, yang masuk dalam kategori "Type II Error". Selain itu, 34 pasien non-stroke salah diprediksi sebagai stroke, yang termasuk dalam kategori "Type I Error".

3.5 Analisis Hasil

Setelah melakukan pengujian model dari skenario pertama hingga keenam pada *dataset stroke prediction*, langkah berikutnya adalah menganalisis hasil dengan memperbandingkan kinerja klasifikasi. Bagian ini akan membahas akurasi, presisi, recall, dan spesifisitas dari setiap skenario pengujian. Selanjutnya, nilai-nilai tersebut akan dibandingkan untuk mencari hasil optimal. Perbandingan hasil pengujian ini diharapkan dapat memberikan solusi terhadap permasalahan yang telah disampaikan sebelumnya.

Tabel 8. Perbandingan Hasil Pengujian 6 Skenario

Uji Validasi	Algoritma Decision Tree	Algoritma Naïve Bayes	Algoritma Decision Tree + Bagging	Algoritma Decision Tree + Adaboost	Algoritma Naïve Bayes + Bagging	Algoritma Naïve Bayes + Adaboost
Akurasi	93,07%	87,14%	96,91%	96,58%	87,22%	94,56%
Presisi	17,39%	15,77%	98,92%	78,46%	15,99%	12,82%
Recall/ Sensitivitas	11,24%	37,75%	36,95%	40,96%	38,15%	2,01%
Spesifisitas	97,26%	89,67%	99,98%	99,42%	89,73%	99,30%

Berdasarkan tabel 8 diatas, hasil perbandingan pengujian pada *Dataset Stroke Prediction* menggunakan algoritma decision tree mendapatkan nilai akurasi, presisi dan spesifisitas yang tinggi dibanding dengan algoritma naïve bayes. Namun, untuk nilai sensitivitas lebih tinggi naïve bayes daripada decision tree. Kemudian pada skenario ketiga yaitu kombinasi algoritma decision tree dan teknik bagging menghasilkan nilai akurasi sebesar 96,91% dimana terdapat peningkatan sebesar 3,84% dari pengujian algoritma decision tree sebelum dikombinasi dengan metode bagging. Selain itu terjadi peningkatan yang sangat signifikan sebesar 81,53% pada nilai presisi. Untuk nilai sensitivitas juga mengalami peningkatan sebesar 25,71% begitupun dengan spesifisitas yang meningkat 2,72%. Pada skenario keempat yaitu kombinasi algoritma decision tree dengan metode adaboost menghasilkan nilai akurasi yang lebih tinggi daripada hanya menggunakan algoritma decision tree saja. Kenaikan akurasi sebesar 3,51% dari 93,07% menjadi 96,58%. Namun, jika dibandingkan dengan skenario ketiga yaitu kombinasi decision tree dan bagging maka nilai akurasi, presisi dan spesifisitas lebih kecil tetapi nilai sensitivitasnya lebih besar. Dapat disimpulkan bahwa teknik bagging memberikan dampak yang lebih besar dalam meningkatkan kinerja algoritma decision tree jika dibandingkan metode adaboost. Selanjutnya, pada skenario kelima yaitu kombinasi algoritma naïve bayes dengan teknik bagging menghasilkan nilai akurasi, presisi, sensitivitas dan spesifisitas yang lebih tinggi daripada hanya menggunakan algoritma naïve bayes saja. Namun, kenaikannya tidak begitu signifikan, rata – rata kenaikannya dibawah 1%. Sehingga dapat disimpulkan teknik bagging tidak terlalu signifikan jika di kombinasi dengan algoritma naïve bayes. Sementara pada skenario keenam yaitu kombinasi algoritma naïve bayes dengan metode adaboost menghasilkan nilai akurasi dan spesifisitas yang tinggi daripada hanya menggunakan algoritma naïve bayes saja. Tetapi, dalam hal presisi dan sensitivitas, algoritma naïve bayes menunjukkan peningkatan nilai yang lebih tinggi setelah dikombinasikan dengan metode adaboost. Namun, jika dibandingkan dengan skenario kelima yaitu kombinasi naïve bayes dan teknik bagging, maka nilai akurasi dan spesifisitas lebih tinggi

menggunakan metode adaboost. Tetapi, untuk nilai presisi dan sensitivitas lebih tinggi menggunakan teknik bagging daripada adaboost. Meskipun keduanya sama-sama memberikan pengaruh terhadap kinerja algoritma naïve bayes, namun peningkatan yang diperoleh dari metode bagging masih kurang signifikan dibandingkan menggunakan metode adaboost. Dari hasil pengujian dan analisis secara keseluruhan, dapat disimpulkan bahwa teknik bagging dan adaboost memiliki dampak yang cukup besar terhadap peningkatan kinerja algoritma decision tree, yakni melebihi 3,5%. Sedangkan pada algoritma naïve bayes hanya metode adaboost yang memberikan pengaruh yang cukup besar terhadap peningkatan kinerja algoritma naïve bayes yaitu mencapai 7,42%. Sementara itu, teknik bagging tidak memberikan peningkatan yang signifikan terhadap kinerja algoritma Naïve Bayes, dengan peningkatan yang kurang dari 1%. Oleh karena itu, dilihat dari keenam skenario pengujian tersebut penerapan teknik bagging pada algoritma decision tree terbukti lebih unggul dibandingkan metode adaboost. Tetapi, pada algoritma naïve bayes penerapan metode adaboost terbukti lebih unggul dibandingkan metode bagging.

4. KESIMPULAN

1. Kombinasi teknik bagging dan adaboost pada algoritma decision tree dan naïve bayes bertujuan untuk mengatasi ketidakseimbangan kelas dan meningkatkan kinerja klasifikasi. Penggunaan teknik bagging dan adaboost memberikan dampak yang signifikan dalam meningkatkan akurasi, presisi, sensitivitas dan spesifisitas algoritma decision tree. Namun, metode Bagging tidak berpengaruh secara signifikan terhadap peningkatan kinerja klasifikasi algoritma Naïve Bayes. Di sisi lain, metode Adaboost menunjukkan pengaruh yang sangat signifikan terhadap kinerja algoritma Naïve Bayes, dengan peningkatan akurasi sebesar 7,42%.
2. Berdasarkan hasil perbandingan enam skenario pengujian, algoritma decision tree mencapai akurasi sebesar 93,07%, presisi 17,39%, spesifisitas 97,26%, dan sensitivitas 11,24%. Setelah menerapkan metode bagging, akurasi meningkat menjadi 96,91%, presisi 98,92%, spesifisitas 99,98%, dan sensitivitas 36,95%. Sementara itu, setelah menggunakan metode adaboost, akurasi menjadi 96,58% dengan presisi 78,46%, spesifisitas 99,42%, dan sensitivitas 40,96%. Algoritma naïve bayes menghasilkan akurasi 87,14%, presisi 15,77%, spesifisitas 89,67%, dan sensitivitas 37,75%. Setelah diterapkan metode bagging, akurasi meningkat menjadi 87,22%, presisi 15,99%, spesifisitas 89,73%, dan sensitivitas 38,15%. Setelah menggunakan metode adaboost, akurasi naik menjadi 94,56% dengan presisi 12,82%, spesifisitas 99,30%, dan sensitivitas 2,01%. Komparasi antara teknik Bagging dan Adaboost pada algoritma decision tree dan naïve bayes menunjukkan peningkatan signifikan dalam kinerja klasifikasi. Akurasi algoritma decision tree meningkat sebesar 3,84% dan 3,51% setelah menerapkan Bagging dan Adaboost. Sementara itu, akurasi algoritma naïve bayes meningkat sebanyak 7,42% setelah menggunakan Bagging dan 0,08% setelah menggunakan Adaboost. Metode Adaboost terbukti signifikan dalam meningkatkan kinerja algoritma naïve bayes, sedangkan Bagging tidak memberikan dampak yang signifikan pada algoritma tersebut.
3. Berdasarkan nilai akurasi dari keenam skenario pengujian, kombinasi algoritma decision tree dengan bagging mendapatkan akurasi tertinggi, mencapai 96,91%. Selanjutnya, kombinasi algoritma decision tree dengan Adaboost mencapai 96,58%, sedangkan kombinasi algoritma Naïve Bayes dengan Adaboost mencapai 94,56%. Algoritma decision tree tanpa metode tambahan mencapai akurasi sebesar 93,07%, kombinasi algoritma Naïve Bayes dengan Bagging mencapai 87,22%. Sedangkan akurasi terendah yaitu algoritma Naïve Bayes dengan nilai 87,14%.

5. SARAN

1. Memperluas penelitian dengan menggunakan algoritma klasifikasi lainnya guna meningkatkan akurasi prediksi.
2. Untuk penelitian selanjutnya bisa mencoba menerapkan teknik ensemble lainnya seperti Random Forest atau Gradient Boosting untuk mengoptimalkan kinerja klasifikasi. Setelah itu bandingkan apakah ada peningkatan akurasi atau justru tidak ada peningkatan sama sekali.
3. Penggunaan teknik sampling yaitu under sampling atau over sampling seperti SMOTE dan teknik sampling yang lain, untuk mengatasi masalah ketidakseimbangan data. Karena *dataset stroke prediction* merupakan dataset yang memiliki kelas data tidak seimbang.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah berperan dalam memberi dukungan **t**erhadap penelitian ini.

DAFTAR PUSTAKA

[1] B. Imran, E. Wahyudi, A. Subki, S. Salman, and A. Yani, "Classification of stroke patients using data mining with adaboost, decision tree and random forest models," *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 218–228, 2022, doi: 10.33096/ilkom.v14i3.1328.218-228.

[2] M. K. Rekha and I. P. Kumar, "Brain Stroke Prediction Using Random Forest And Adaboost Algorithm," no. 6, pp. 84–94, 2023.

[3] E. F. Santika, "No Title," *katadata*. https://databoks.katadata.co.id/datapublish/2023/02/07/stroke-dan-tbc-masuk-dalam-10-penyakit-penyebab-kematian-tertinggi-di-indonesia (accessed Nov. 21, 2023).

[4] N. Qur, "No Title," *detikJatim*. https://www.detik.com/jatim/berita/d-7008701/rtpa-upaya-indonesia-cegah-kematian-dan-kurangi-kecacatan-akibat-stroke (accessed Nov. 21, 2023).

[5] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *J. SAINTEKOM*, vol. 13, no. 1, pp. 42–54, 2023, doi: 10.33020/saintekom.v13i1.352.

[6] V. N. Sari, L. Y. Astri, and E. Rasywir, "Analisis dan Penerapan Algoritma Naive Bayes untuk Evaluasi Kinerja Karyawan pada PT. Pelita Wira Sejahtera," *J. Ilm. Mhs. Tek. Inform.*, vol. 2, no. 1, pp. 53–68, 2020.

[7] Y. T. U. Heni Sulistiani, "Penerapan Algoritma Klasifikasi Sebagai Pendukung Keputusan Pemberian Beasiswa Mahasiswa," *Snti*, no. October 2018, pp. 300–305, 2018.

[8] A. P. Permana, K. Ainiyah, and K. F. H. Holle, "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 3, pp. 178–188, 2021, doi: 10.14421/jiska.2021.6.3.178-188.

[9] A. Ridwan, "Penerapan Teknik Bagging Pada Algoritma Naive Bayes," *J. Bisnis Digit. Dan Sist. Inf.*, vol. 1, no. 1, pp. 63–70, 2020, [Online]. Available: https://ejr.stikesmuhkudus.ac.id/index.php/BIDISFO/article/view/914

[10] G. P. Kawani, "Implementasi Naive Bayes," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 73–81, 2019, doi: 10.20895/inista.v1i2.73.

[11] I. Yulianti, R. Amegia Saputra, M. Sukrisno Mardiyanto, and A. Rahmawati, "Accuracy Optimization of C4.5 Algorithm Based on Particle Swarm Optimization with Bagging Technique on Prediction of Chronic Kidney Disease," *Techno.COM*, vol. 19, no. 4, pp. 411–421, 2020, [Online]. Available: https://archive.ics.uci.edu/ml/

[12] A. Nur Rais and Warjiyono, "Optimasi Akurasi Klasifikasi Pada Prediksi Smokte

Detection dengan Menggunakan Algoritma Adaboost," *J. Sist. Komput. dan Inform.*, vol. 4, no. 2, pp. 343–348, 2022, doi: 10.30865/json.v4i2.5154.

[13] N. D. Saputri, K. Khalid, and D. Rolliawati, "Komparasi Penerapan Metode Bagging dan Adaboost pada Algoritma C4 . 5 untuk Prediksi Penyakit Stroke," *Sist. J. Sist. Inf.*, vol. 11, no. September, pp. 567–577, 2022.

[14] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.

[15] K. L. Kohsasih and Z. Situmorang, "Comparative Analysis of C4.5 and Naïve Bayes Algorithms in Predicting Cerebrovascular Disease," *J. Inform.*, vol. 9, no. 1, pp. 13–17, 2022.

[16] T. Setiyorini and R. S. Wahono, "Penerapan Metode Bagging Untuk Mengurangi Data Noise Pada Neural Network Untuk Estimasi Kuat Tekan Beton," *J. Intell. Syst.*, vol. 1, no. 1, pp. 36–41, 2015.

[17] Wasilah; Halimah, "Analysis of Graduate Success Patterns Based on Association Rule Mining to Increase the Achievement of the Performance Index of Higher Education Graduates," *Int. J. Artif. Intelegence Res.*, vol. 7, no. 1, pp. 1–8, 2023, [Online]. Available: https://ijair.id/index.php/ijair/article/view/1095/pdf