Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

Komparasi Perbandingan Algoritma C4.5, Naive Bayes, K-Nearest Neighbor, Random Forest Untuk Prediksi Faktor Penyebab Penyakit Diabetes

Muhammad Bagus Fadli*, Iwan Purnama, Rohani

Fakultas Sains dan Teknologi, Program Teknologi Informasi, Universitas Labuhanbatu, Rantauprapat, Indonesia
Email : [1,*]muhammadbagusfadli257@gmail.com, [2]iwanpurnama2014@ulb.ac.id, [3]pasariburohani@gmail.com
Email Penulis: muhammadbagusfadli257@gmail.com
Submitted: **11/11/2025**; Accepted: **31/12/2025**; Published**: 31/12/2025**

Abstrak−Diabetes merupakan penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah dan dapat menyebabkan berbagai komplikasi serius serta berkontribusi terhadap tingginya angka kematian di dunia. Permasalahan utama dalam penanganan penyakit diabetes adalah perlunya klasifikasi status pasien yang akurat berdasarkan data pemeriksaan laboratorium agar dapat dilakukan penanganan yang tepat. Penelitian ini bertujuan untuk mengkomparasikan kinerja algoritma C4.5, Naive Bayes, K-Nearest Neighbor (KNN), dan Random Forest dalam mengklasifikasikan data pasien diabetes. Dataset yang digunakan bersumber dari *Electronic Health Records* (EHRs) dengan subjek penelitian dari Rumah Sakit Umum Daerah Rantauprapat, berjumlah 10.000 data yang terdiri dari delapan atribut dan satu atribut kelas, dengan 859 data pasien diabetes dan 9.141 data pasien non-diabetes. Metode penelitian dilakukan dengan membagi data menjadi data training dan data testing menggunakan rasio 90:10, 80:20, dan 70:30. Evaluasi kinerja model menggunakan parameter accuracy dan *Receiver Operating Characteristic* (ROC) dengan nilai *Area Under Curve* (AUC). Hasil penelitian menunjukkan bahwa algoritma C4.5 dan Random Forest menghasilkan nilai accuracy yang lebih tinggi dibandingkan Naive Bayes dan KNN, terutama pada rasio data training 90%:10% dan 70%:30%. Berdasarkan evaluasi ROC, algoritma Random Forest memperoleh nilai AUC tertinggi pada rasio 70%:30% sebesar 0,972 dan 80%:20% sebesar 0,970. Berdasarkan hasil pengujian tersebut, dapat disimpulkan bahwa algoritma C4.5 dan Random Forest memiliki kinerja yang relatif lebih baik dan hampir setara dalam klasifikasi penyakit diabetes berdasarkan nilai accuracy dan AUC.

Kata kunci : Diabetes; Decision Tree; C4.5; Naive Bayes; K-Nearest Neighbor; Random Forest

Abstract−Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels and can cause various serious complications and contribute to high mortality rates worldwide. The main problem in managing diabetes is the need for accurate patient status classification based on laboratory test data so that appropriate treatment can be carried out. This study aims to compare the performance of the C4.5 algorithm, Naive Bayes, K-Nearest Neighbor (KNN), and Random Forest in classifying diabetes patient data. The dataset used was sourced from Electronic Health Records (EHRs) with research subjects from Rantauprapat Regional General Hospital, totaling 10,000 data consisting of eight attributes and one class attribute, with 859 diabetes patient data and 9,141 non-diabetes patient data. The research method was carried out by dividing the data into training data and testing data using a ratio of 90:10, 80:20, and 70:30. Evaluation of model performance used accuracy parameters and Receiver Operating Characteristic (ROC) with Area Under Curve (AUC) values. The results showed that the C4.5 and Random Forest algorithms produced higher accuracy values than Naive Bayes and KNN, especially at training data ratios of 90%:10% and 70%:30%. Based on the ROC evaluation, the Random Forest algorithm obtained the highest AUC values at the 70%:30% ratio of 0.972 and 80%:20% of 0.970. Based on these test results, it can be concluded that the C4.5 and Random Forest algorithms have relatively better performance and are almost equivalent in classifying diabetes based on accuracy and AUC values.

Keywords: Diabetes; Decision Tree; C4.5; Naive Bayes; K-Nearest Neighbor; Random Forest

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi yang pesat pada era digital saat ini telah membawa dampak signifikan terhadap berbagai bidang kehidupan manusia, termasuk bidang kesehatan. Di bidang medis, teknologi telah berperan penting dalam membantu proses diagnosis, pengobatan, serta prediksi penyakit secara lebih cepat dan akurat. Salah satu bidang ilmu yang banyak dimanfaatkan dalam dunia kesehatan adalah data mining dan *machine learning*, yaitu teknik pengolahan data secara cerdas untuk menemukan pola dan pengetahuan baru dari sekumpulan data yang besar (*big data*). Salah satu contoh penggabungan dunia teknologi dengan dunia medis adalah dengan cara memprediksi diabetes. Ada banyak cara untuk memprediksi diabetes, salah satunya adalah menggunakan ilmu yang bernama data mining. Data mining salah satu bidang ilmu dalam komputer yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar, yang nantinya kumpulan informasi data tersebut akan diolah dan akan menghasilkan informasi yang lebih akurat atau menampilkan informasi yang terbaik [1]. Sedangkan menurut Rabel Duta Apyuma, (2025), *data mining* adalah proses menemukan pola yang belum diketahui sebelumnya atau informasi yang berguna dari data mentah [2] Definisi ini mencerminkan konsep umum bahwa data mining melibatkan penggunaan algoritma dan teknik statistik untuk menggali pengetahuan yang tersembunyi atau pola yang berharga dari data besar.

Penelitian yang dilakukan oleh oleh Minyechil, (2023), Penelitian ini membahas analisis dan prediksi penyakit diabetes dengan membandingkan beberapa algoritma machine learning, yaitu Random Forest, K-Nearest Neighbor (KNN), Naive Bayes, dan Decision Tree J48 (C4.5). Dataset yang digunakan berasal dari Pima Indian Diabetes Dataset yang berisi data medis pasien. Penelitian ini menekankan pentingnya pemilihan algoritma klasifikasi yang

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

tepat untuk meningkatkan akurasi prediksi diabetes. Hasil pengujian menunjukkan bahwa Random Forest memiliki tingkat akurasi tertinggi dibandingkan algoritma lainnya, diikuti oleh *C4.5*. *Naive Bayes* dan *KNN* memiliki performa yang lebih rendah, terutama ketika data memiliki korelasi antar atribut. Penelitian ini relevan karena menunjukkan keunggulan algoritma berbasis pohon keputusan dan ensemble dalam klasifikasi diabetes [3].

Penelitian serupa dilakukan oleh Faruque dan rekan, (2019), melakukan perbandingan beberapa algoritma klasifikasi, antara lain *C4.5, Naive Bayes, KNN,* dan *Support Vector Machine* (*SVM*), untuk memprediksi penyakit diabetes melitus. Penelitian ini menggunakan dataset klinis pasien diabetes dan mengevaluasi model menggunakan metrik akurasi, presisi, *recall*, dan *F-measure*. Hasil penelitian menunjukkan bahwa *algoritma C4.5* mampu menghasilkan performa yang stabil dan mudah diinterpretasikan oleh tenaga medis. Naive Bayes menunjukkan kinerja yang baik pada data berskala besar, namun kurang optimal ketika atribut saling bergantung. Penelitian ini mendukung penggunaan *C4.5* sebagai metode klasifikasi yang efektif dalam sistem pendukung keputusan medis [4].

Muhammad Rousydi Hunafa dan Arif Hermawan melakukan perbandingan algoritma *Naive Bayes* dan *K-Nearest Neighbor* pada *Imbalace Class Data Set* Penyakit Diabetes. Hasil penelitian menunjukkan bahwa *Naive Bayes* dengan teknik *SMOTE* menunjukkan performa terbaik dengan akurasi 71.66%, diikuti oleh *Naive Bayes* tanpa *SMOTE* (76.03%), dan *KNN* dengan *SMOTE* (80.47%). Meskipun *KNN* tanpa *SMOTE* memiliki akurasi tertinggi (83.02%), *Naive Bayes* dengan *SMOTE* menunjukkan keseimbangan yang lebih baik antara akurasi, presisi, dan *recall*. Penggunaan teknik SMOTE meningkatkan performa *Naive Bayes* dengan peningkatan presisi dan recall, menunjukkan kemampuannya dalam mengatasi ketidakseimbangan kelas pada dataset diabetes. Studi ini memberikan wawasan tentang pemilihan algoritma terbaik dan teknik penanganan ketidakseimbangan kelas yang efektif dalam memprediksi penyakit diabetes pada dataset yang tidak seimbang [5].

Penelitian ini membandingkan algoritma C4.5 dan Naive Bayes dalam memprediksi penyakit diabetes menggunakan dataset dari rumah sakit di Indonesia ini dilakukan oleh Pratama dan kurniawan, (2023). Proses penelitian meliputi tahap preprocessing data, seleksi atribut, pembagian data training dan testing, serta evaluasi model menggunakan confusion matrix. Hasil penelitian menunjukkan bahwa algoritma C4.5 menghasilkan nilai akurasi yang lebih tinggi dibandingkan Naive Bayes karena kemampuannya menangani data numerik dan kategorikal secara bersamaan. Penelitian ini relevan dengan konteks penelitian di Indonesia dan menunjukkan bahwa C4.5 lebih sesuai untuk data rekam medis rumah sakit [6].

Safitri dan Hidayati, (2022), melakukan perbandingan antara algoritma K-Nearest Neighbor dan Naive Bayes dalam klasifikasi risiko penyakit diabetes. Penelitian ini berfokus pada pengaruh pemilihan nilai parameter K pada algoritma KNN dan asumsi independensi pada Naive Bayes. Hasil penelitian menunjukkan bahwa KNN sensitif terhadap pemilihan nilai K dan skala data, sehingga memerlukan normalisasi data agar menghasilkan performa optimal. Sementara itu, Naive Bayes lebih sederhana dan cepat, namun akurasinya lebih rendah dibandingkan KNN pada dataset tertentu. Penelitian ini memberikan gambaran kelebihan dan kelemahan masing-masing algoritma [7].

Berdasarkan tujuh penelitian terdahulu tersebut, dapat disimpulkan bahwa *algoritma C4.5* dan *Random Forest* secara konsisten menunjukkan kinerja yang lebih baik dibandingkan *Naive Bayes* dan *K-Nearest Neighbor* dalam klasifikasi penyakit diabetes. Random Forest unggul dari sisi akurasi dan AUC, sedangkan C4.5 unggul dari sisi interpretabilitas model. Hal ini menjadi dasar kuat bagi penelitian Anda untuk melakukan komparasi keempat algoritma tersebut menggunakan dataset *EHRs* dari RSUD Rantauprapat.

Penggunaan data mining dengan algoritma seperti C4.5, *Naive bayes*, *K-Nearest Neighbor*, *Deep Learning*, *Artificial Neural Network* (*ANN*), *Support Vector Machine* dan *Generalized Linear Model* (*GLM*). Namun dari algoritma tersebut yang paling banyak digunakan penelitian adalah algoritma C4.5 dan algoritma *Naive bayes*, mereka berfokus untuk mencari nilai *accuracy* dari algoritma yang mereka gunakan serta tidak menggunakan validasi data. Padahal jika menggunakan pengujian *accuracy* pada masing-masing algoritma akan lebih akurat algoritma mana yang terbaik untuk menghasilkan pola. Sebagai pilihan untuk diagnosa penyakit diabetes dapat menjadi alternatif pilihan yang tepat, tetapi sampai saat ini belum diketahui algoritma yang paling akurat dalam memprediksi penyakit diabetes.

Berdasarkan latar belakang diatas maka Tujuan dari penelitian ini adalah untuk mengkomparasikan algoritma *C4.5, Naive bayes, K-Nearest Neighbor* dan *Random Forest* dalam penentuan klasifikasi data pasien diabetes RSUD Rantauprapat.

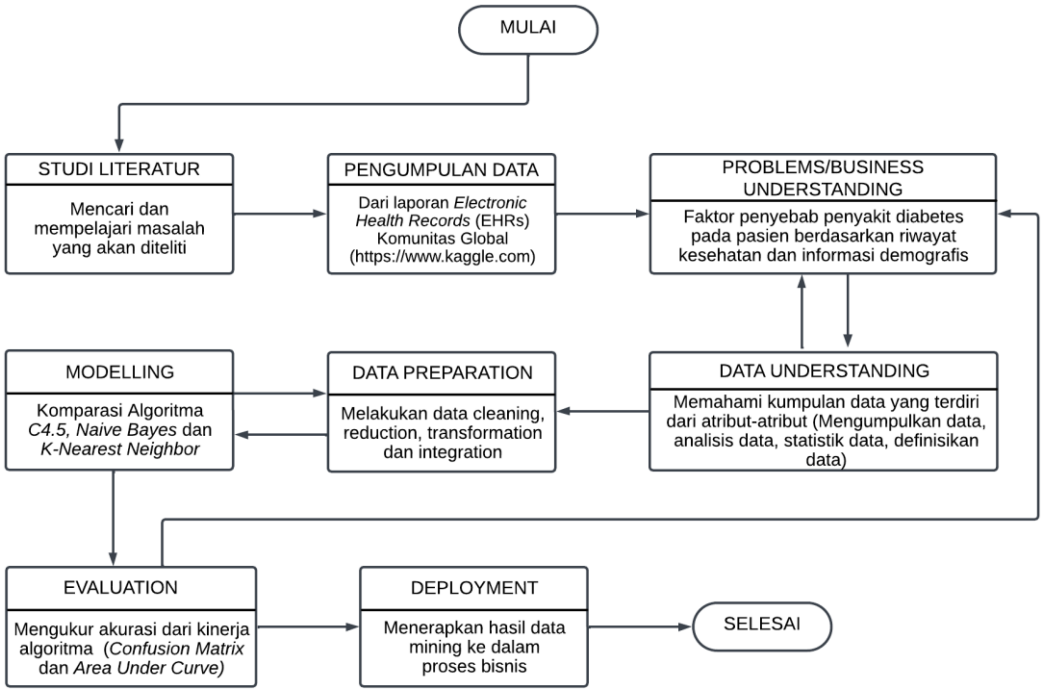
2. METODE PENELITIAN

Metode dasar yang digunakan dalam penelitian ini adalah metode kuantitatif. Tujuannya yaitu membangun model klasifikasi dan mengukur kinerja pada masing-masing model untuk mendapatkan kinerja yang terbaik dari model yang dihasilkan dengan algoritma C4.5, *Naive bayes*, *K-Nearest Neighbor* (*K-NN*) dan *Random Forest*. *C4.5* merupakan algoritma klasifikasi berbasis *decision tree* yang mengembangkan *algoritma ID3*. *Algoritma* ini membangun pohon keputusan berdasarkan nilai *gain ratio* untuk memilih atribut terbaik sebagai pemisah data. *C4.5* mampu menangani data numerik dan kategorik, serta mengatasi *missing value*. Hasil klasifikasi mudah dipahami karena berbentuk aturan keputusan. *Naive Bayes* adalah algoritma klasifikasi berbasis probabilistik yang menggunakan *Teorema Bayes* dengan asumsi bahwa setiap atribut bersifat independen. Algoritma ini memiliki proses komputasi yang sederhana dan cepat, serta efektif digunakan pada dataset berukuran besar. Meskipun asumsi independensinya sederhana, *Naive Bayes* sering menghasilkan performa yang cukup baik. *KNN* merupakan algoritma *lazy learning* yang melakukan klasifikasi

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

berdasarkan kedekatan jarak data uji dengan sejumlah data latih terdekat (k). Penentuan kelas dilakukan berdasarkan mayoritas kelas tetangga terdekat. *KNN* mudah diimplementasikan, namun sensitif terhadap pemilihan nilai k dan skala data. *Random Forest* adalah algoritma *ensemble learning* yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi klasifikasi. Setiap pohon dibangun dari sampel data dan atribut yang dipilih secara acak. Algoritma ini mampu mengurangi *overfitting* dan memberikan performa yang baik pada data kompleks, meskipun membutuhkan sumber daya komputasi lebih besar.

Siklus hidup dalam gambaran umum untuk proses penelitian yang dilakukan oleh peneliti sebagai berikut :



Gambar 1. Tahapan Metode Penelitian

Gambar 1 menggambarkan tahapan proses penelitian data mining dalam studi dengan subjek primer Rumah Sakit Umum Daerah (RSUD) Rantauprapat. Langkah-langkahnya mengikuti standar CRISP-DM (Cross-Industry Standard Process for Data Mining) yang terdiri dari beberapa fase utama: mulai dari pemahaman masalah hingga penerapan hasil.

a. Mulai
   Tahap pertama menandai awal proses penelitian. Pada tahap ini peneliti menentukan fokus dan arah penelitian, yaitu menganalisis serta membandingkan performa beberapa algoritma *machine learning* dalam memprediksi faktor penyebab penyakit diabetes berdasarkan data kesehatan pasien. [8].
b. Studi Literatur
   Sumber dari studi literatur yang digunakan oleh peneliti adalah menggunakan berbagai jurnal international, jurnal yang terindeks dalam sinta 2 sampai sinta 4, repositori dataset publik dan juga dari buku tesk dan literatur yang digunakan di akademik. Dari data tersebut dapat diperoleh temuan penting yang berkaitan dengan judul yang telah dibuat oleh peneliti, dengan melakukan sebuah pengembangan berdasarkan subjek penelitian primer dari Rumah Sakit Umum Daerah (RSUD) Rantauprapat [9].
   Secara garis besar ada dua tipe *machine learning*, yaitu *Supervised Learning* dan *Unsupervised Learning*. *Supervised learning* mengacu pada machine learning dimana data yang digunakan untuk belajar sudah diberi label *output* yang harus dikeluarkan mesin, sedangkan *Unsupervised learning* sebaliknya mengacu pada *machine learning* yang belajar dari data yang tidak diberi label *output* [10].
c. Pengumpulan Data
   Pada tahap ini peneliti mengumpulkan data yang akan digunakan sebagai bahan analisis. Sumber data berasal dari *Electronic Health Records* (*EHRs*). Dataset ini berisikan tentang usia pasien, jenis kelamin, kadar glukosa, tekanan darah, indeks masa tubuh (IMT), insulin, tebal kulit dan riwayat keluarga dan status diabetes.
d. *Problems/Business Understanding*
   Tahapan ini bertujuan untuk memahami permasalahan utama yang ingin diselesaikan. Dalam penelitian ini, masalah utamanya adalah "Faktor apa saja yang paling berpengaruh terhadap timbulnya penyakit diabetes, dan algoritma mana yang paling akurat dalam melakukan prediksi tersebut."
e. Data *Understanding*
   Tahapan ini bertujuan untuk memahami struktur, karakteristik, dan isi dataset yang dikumpulkan. Langkah-langkah tersebut adalah melakukan eksplorasi data (*data exploration*), mengidentifikasi atribut atau variabel

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

penting (misalnya kadar glukosa, BMI, usia, tekanan darah) dan menganalisis distribusi data dan mendeteksi *missing values* atau *outlier*.

f. Data *Preparation*
Tahapan ini dikenal juga dengan istilah *preprocessing data*, yaitu proses menyiapkan data agar siap digunakan dalam tahap modelling. Beberapa kegiatan utama di tahap ini meliputi *data Cleaning* yaitu menghapus data kosong, duplikat, atau nilai tidak valid, *data Reduction* yaitu menghapus atribut yang tidak relevan agar data lebih efisien, *data Transformation* yaitu melakukan normalisasi atau standarisasi agar semua atribut memiliki skala yang sama dan *data Integration* merupakan menggabungkan data dari beberapa sumber jika diperlukan [11].

g. *Modelling*
Pada tahap ini dilakukan proses pembangunan model prediksi menggunakan beberapa algoritma *machine learning* [12]. Empat algoritma yang digunakan dalam penelitian ini adalah *C4.5* (*Decision Tree*) yang membentuk struktur pohon keputusan berdasarkan atribut yang paling berpengaruh, *naive bayes* yang menggunakan metode probabilistik untuk menghitung kemungkinan seseorang menderita diabetes berdasarkan atribut kesehatan, *k-nearest neighbor* (*KNN*) yang mengklasifikasikan data baru berdasarkan kedekatannya dengan data yang sudah ada dan juga *random forest* yang menggabungkan beberapa pohon keputusan (C4.5) untuk meningkatkan akurasi dan mengurangi overfitting [13]

h. *Evaluation*
Setelah model terbentuk, langkah selanjutnya adalah mengukur performa masing-masing algoritma**.** Evaluasi dilakukan dengan menggunakan beberapa metrik, antara lain [14] :
1. *Accuracy* : tingkat ketepatan model dalam mengklasifikasikan data.
2. *Confusion Matrix* : tabel yang menunjukkan perbandingan antara hasil prediksi dengan data aktual (*True Positive, False Positive, True Negative, False Negative).*
3. *Precision, Recall,* dan *F1-Score* : untuk mengukur seberapa baik model mengenali kasus positif diabetes.
4. *Area Under Curve* (*AUC*) : mengukur kemampuan model dalam membedakan kelas positif dan negatif.

i. *Deployment*
Tahapan ini adalah penerapan hasil penelitian dalam bentuk sistem nyata atau simulasi. Misalnya, hasil komparasi algoritma yang terbaik (misalnya *Random Forest*) dapat diimplementasikan ke dalam sistem pendukung keputusan (*Decision Support System*) atau aplikasi prediksi diabetes berbasis *web* [14].

j. Selesai
Merupakan akhir dari seluruh proses penelitian. Tahap ini menandakan bahwa semua langkah-mulai dari studi literatur hingga implementasi hasil-telah dilakukan, dan peneliti dapat menarik kesimpulan akhir serta memberikan saran untuk penelitian selanjutnya [15].

3. HASIL DAN PEMBAHASAN

Penelitian ini menerapkan model *Cross-Standard Industry for Data Mining* (*CRISP-DM*) untuk mengkomparasikan algoritma C4.5, Naive Bayes, K-Nearest Neighbor, dan Random Forest dalam prediksi faktor penyebab penyakit diabetes berdasarkan data *Electronic Health Records* (EHRs) RSUD Rantauprapat. Penerapan CRISP-DM mampu memberikan alur penelitian yang sistematis mulai dari pemahaman masalah hingga evaluasi model, sehingga proses analisis data dapat dilakukan secara terstruktur dan efektif, [8].

Data understanding, data yang didapat berasal dari laporan *Electronic Health Records* (*EHRs*) adalah sumber data utama untuk kumpulan data prediksi diabetes yang melibatkan pengumpulan data medis dan demografi dari pasien yang telah didiagnosis atau berisiko terkena diabetes. Jumlah data yang digunakan sebanyak 10.000 *record*, memiliki delapan atribut dan satu atribut dengan status pasien sebagai label (*class*) yang menyatakan 859 data pasien yang menderita penyakit diabetes, 9141 data pasien yang tidak menderita penyakit diabetes. Dapat dilihat pada Tabel 1.

Dalam penelitian ini, eksperimen dilaksanakan dengan tujuan untuk menentukan tingkat akurasi terbaik di antara *algoritma C4.5, Naive Bayes, K-Nearest Neighbor* (*K-NN*), dan *Random Forest*. Keempat algoritma ini dibandingkan untuk menemukan yang paling efektif. Setelah pembuatan model, pengujian dilakukan dengan menggunakan 10-*fold cross validation*. Perbandingan antara data training dan data testing yang digunakan adalah sebagai berikut : 90% data *training* dan 10% data *testing*, 80% data *training* dan 20% data *testing*, serta 70% data *training* dan 30% data *testing*.

Tabel 1. Sampel Data Training

Gender	Age	Hyper Tention	Heart_disease	Smoking_History	BMI	HbA1c_Level	20	Diabetes
Female	80	0	0	Never	25.2	6.6	140	0
Female	54	0	0	Not Info	27.3	6.6	80	0
Male	28	0	0	Never	27.3	5.7	158	0
Female	36	0	0	Not Current	25.3	5	155	0
Male	76	1	1	Not Current	20.1	4.8	155	0
Female	20	0	0	Never	27.3	6.6	85	0

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

Gender	Age	Hyper Tention	Heart_disease	Smoking_History	BMI	HbA1c_Level	20	Diabetes
Female	44	0	0	Never	19.3	6.5	200	1
Female	79	0	0	Not Info	23.9	5.7	85	0
Male	42	0	0	Never	33.6	4.8	145	0
Female	32	0	0	Never	27.3	5	100	0
Female	53	0	0	Never	27.3	6.1	85	0
Female	54	0	0	Former	54.7	6	100	0
Male	78	0	0	Former	36.1	5	130	0
Female	67	0	0	Former	25.7	5.8	200	0
Female	76	0	0	Never	27.3	5	160	0
Female	78	0	0	Not Info	27.3	6.6	126	0
Male	15	0	0	Not Info	30.4	6.1	200	0
Female	42	0	0	Never	24.5	5.7	158	0
Female	42	0	0	Not Info	27.3	5.7	80	0

Tabel 1 memberikan keterangan bahwa Tabel ini menyajikan data latih penting untuk model klasifikasi, misalnya model *C4.5, Naive Bayes, K-Nearest Neighbor* (*KNN*), atau *Random Forest***.** Model akan belajar dari data ini untuk menemukan pola hubungan antar variabel seperti usia, tekanan darah, riwayat merokok, dan kadar HbA1c terhadap kemungkinan seseorang mengidap diabetes atau tidak [16].

Berikutnya, hasil komparasi algoritma yang telah dilakukan berdasarkan data training dan data testing yang sudah ditemukan.

Tabel 2. Hasil Komparasi Algoritma Yang Telah Dilakukan Berdasarkan Data Training Dan Data Testing

Algoritma	Data *Training*	Data *Testing*	*Accuracy*	*ROC/AUC*
C4.5	90	10	97.21%	0.874
	80	20	97.15%	0.900
	70	30	97.31%	0.732
Naive Bayes	90	10	65.64%	0.953
	80	20	95.56%	0.953
	70	30	95.57%	0.955
K-Nearest Neighboar	90	10	94.91%	0.866
	80	20	94.86%	0.872
	70	30	94.84%	0.868
Random Forest	90	10	97.29%	0.972
	80	20	97.21%	0.970
	70	30	97.33%	0.972

Dari Tabel 2, kita dapat melihat hasil komparasi dari keempat algoritma yang digunakan dalam penelitian ini, yaitu *C4.5, Naive Bayes, K-Nearest Neighbor*, dan *Random Forest*. Komparasi dilakukan berdasarkan pembagian data *training* dan data *testing* dengan perbandingan 90:10, 80:20, dan 70:30. Pada pembagian data *training* sebesar 90% dan data *testing* sebesar 10%, *algoritma Random Forest* dan *C4.5* menunjukkan nilai accuracy tertinggi. *Random Forest* mencapai accuracy 97.29%, sedangkan *C4.5* mencapai 97.21%. Hal ini lebih tinggi dibandingkan dengan *Naive Bayes* yang mencapai 95.64% dan *K-Nearest Neighbor* dengan akurasi terendah 94.91%. Selain itu, nilai *Area Under Curve* (*AUC*) dari *Random Forest* juga menunjukkan performa terbaiknya dibandingkan ketiga algoritma lainnya, yaitu *C4.5, Naive Bayes*, dan *K-Nearest Neighbor*. Untuk pembagian data training 80% dan data testing 20%, Random Forest kembali menunjukkan nilai *accuracy* dan *AUC* yang terbesar. Sedangkan yang terakhir, untuk percobaan perbandingan data *training* 70% data *testing* 30% tidak ada perbedaan dengan percobaan kedua yang dimana menghasilkan nilai *accuracy* dan *AUC* tertinggi adalah *algoritma Random Forest* memiliki nilai *accuracy* sebesar 97.33%, nilai *AUC* sebesar 0.972. Untuk rata-rata keseluruhan percobaan dapat dilihat pada tabel 3 dibawah ini [17]

Tabel 3. Rata-rata Hasil Komparasi Algoritma Yang Telah Dilakukan Berdasarkan Data Training Dan Data Testing

Algoritma	Data *Training*	Data *Testing*	*Accuracy*	*ROC/AUC*
C4.5	90	10	97.21%	0.874
	80	20	97.15%	0.900
	70	30	97.31%	0.732
	90	10	65.64%	0.953
	Rata-rata		97.22%	0.835
Naive Bayes	80	20	95.56%	0.953
	70	30	95.57%	0.955
	90	10	94.91%	0.866
	Rata-rata		95.59%	0.953
	80	20	94.86%	0.872

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

Algoritma	Data *Training*	Data *Testing*	*Accuracy*	*ROC/AUC*
	70	30	94.84%	0.868
K-Nearest Neighboar	90	10	97.29%	0.972
	Rata-rata		94.87%	0.869
	80	20	97.21%	0.970
	70	30	97.33%	0.972
Random Forest	Rata-rata		97.28%	0.971

Dari Tabel 3, rata-rata akurasi dari *algoritma Random Forest* mencapai 97.28%. Ini merupakan rata-rata akurasi tertinggi dibandingkan dengan *algoritma C4.5, Naive Bayes*, dan *K-Nearest Neighbor*. Selain itu, untuk rata-rata nilai *Area Under Curve* (*AUC*), *Random Forest* juga menunjukkan *performa* terbaik dengan nilai *AUC* sebesar 0.971. Model *confusion matrix* akan menghasilkan matriks yang terdiri dari empat bagian : *true positif, false positif, true negatif,* dan *false negatif*. Berikut dibawah ini merupakan hasil *confusion matrix* dari algoritma klasifikasi *C4.5*, *Naive bayes, K-Nearest Neighbor* dan *Random Forest* untuk data *training* 90% data *testing* 10% sebagai *accuracy* yang paling tinggi didapatkan [18].

accuracy: 97.21% +/- 0.69% (micro average: 97.21%)

	true Tidak	true Ya	class precision
pred. Tidak	8218	242	97.14%
pred. Ya	9	531	98.33%
class recall	99.89%	68.69%	

Gambar 2. *Confussion matrix Algoritma C4.5* (Data *training* 90%, Data *testing* 10%).

Penjelasan dari Gambar 2 menunjukan bahwa, diketahui terdapat sebanyak 531 jumlah data yang diprediksi diabetes dan pada kenyataannya memang menderita penyakit diabetes, 8218 data diprediksi tidak diabetes dan pada kenyataannya memang tidak menderita penyakit diabetes, 9 data yang diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes, dan 242 data diprediksi tidak menderita diabetes tetapi kenyataannya diabetes [19]. Berdasarkan Gambar 2 menunjukan bahwa, tingkat *accuracy* dengan menggunakan *algoritma C4.5* untuk perbandingan data *training* dan data *testing* 90% : 10% adalah sebesar 97.21%.

Model *confussion matrix* yang kedua dengan menggunakan algoritma klasifikasi *Naive bayes*, Untuk perbandingan data *training* dan data *testing* 90% : 10% sehingga didapatkan hasil pada gambar 3 sebagai berikut :

accuracy: 95.64% +/- 0.76% (micro average: 95.64%)

	true Tidak	true Ya	class precision
pred. Tidak	8089	254	96.96%
pred. Ya	138	519	79.00%
class recall	98.32%	67.14%	

Gambar 3. *Confussion matrix Algoritma Naive bayes* (Data *training* 90%, Data *testing* 10%).

Pada Gambar 3, menunjukan bahwa, sebanyak 8.089 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 519 data diprediksi diabetes dan pada kenyataannya memang diabetes, 254 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 138 data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes [20]. Berdasarkan gambar 3 menunjukan bahwa, tingkat *accuracy* dengan menggunakan *algoritma Naive Bayes* untuk perbandingan data *training* dan data *testing* 90% : 10% adalah sebesar 95.64%.

Model *confussion matrix* yang ketiga dengan menggunakan *algoritma K-Nearest Neighbor*, Untuk perbandingan data *training* dan data *testing* 90% : 10% sehingga didapatkan hasil pada gambar 4 sebagai berikut :

accuracy: 94.91% +/- 0.65% (micro average: 94.91%)

	true Tidak	true Ya	class precision
pred. Tidak	8171	402	95.31%
pred. Ya	56	371	86.89%
class recall	99.32%	47.99%	

Gambar 4. *Confussion matrix Algoritma K-Nearest Neighbor* (Data *training* 90%, Data *testing* 10%).

Pada Gambar 4 menunjukan bahwa, sebanyak 8.171 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 371 data diprediksi diabetes dan pada kenyataannya memang diabetes, 402 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 56

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes. Berdasarkan gambar 4, menunjukan bahwa, tingkat *accuracy* dengan menggunakan *algoritma K-Nearest Neighbor* untuk perbandingan data *training* dan data testing 90% : 10% adalah sebesar 94.91%.

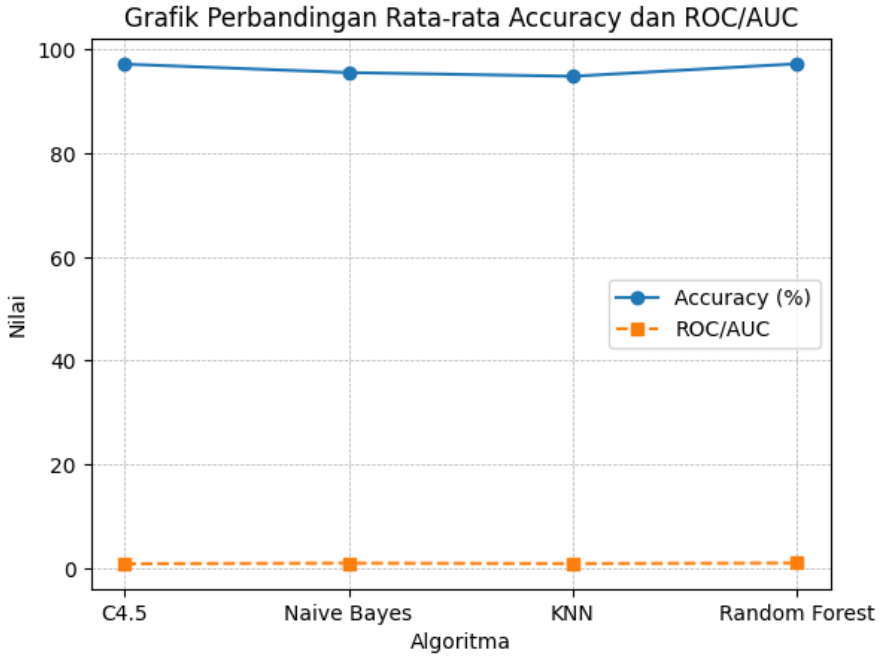
Model *confussion matrix* yang keempat dengan menggunakan *algoritma Random Forest*, Untuk perbandingan data *training* dan data *testing* 90% : 10% sehingga didapatkan hasil pada gambar 5 sebagai berikut :

accuracy: 97.29% +/- 0.35% (micro average: 97.29%)

	true Tidak	true Ya	class precision
pred. Tidak	8223	240	97.16%
pred. Ya	4	533	99.26%
class recall	99.95%	68.95%	

Gambar 5. *Confussion matrix Algoritma Random Forest* (Data *training* 90%, Data *testing* 10%).

Pada Gambar 5, menunjukan bahwa, sebanyak 8.223 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 533 data diprediksi diabetes dan pada kenyataannya memang diabetes, 240 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 4 data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes [21]. Berdasarkan gambar 5, menunjukan bahwa, tingkat *accuracy* dengan menggunakan *algoritma Random Forest* untuk perbandingan data *training* dan data *testing* 90%: 10% yang paling tertinggi adalah sebesar 97.29%.



Gambar 6. Grafik Perbandingan Rata-rata *Accuracy* dan *ROC/AUC*

Berdasarkan Gambar 6, yang menunjukkan perbandingan rata-rata *Accuracy* dan *ROC/AUC* dari algoritma *C4.5, Naive Bayes, K-Nearest Neighbor (KNN),* dan *Random Forest*, dapat diketahui bahwa setiap algoritma memiliki tingkat performa yang berbeda dalam memprediksi faktor penyebab penyakit diabetes. Algoritma *Random Forest* menunjukkan performa paling optimal dibandingkan algoritma lainnya. Hal ini ditunjukkan oleh nilai *accuracy* tertinggi sebesar 97,28**%** dan *ROC/AUC* sebesar 0,971, yang mengindikasikan kemampuan model yang sangat baik dalam mengklasifikasikan data pasien diabetes dan non-diabetes. Tingginya nilai *ROC/AUC* menunjukkan bahwa *Random Forest* memiliki tingkat sensitivitas dan spesifisitas yang seimbang serta stabil dalam menangani data yang kompleks. Algoritma *C4.5* menempati posisi kedua dengan nilai *accuracy* sebesar 97,22**%**, yang relatif mendekati *Random Forest*. Namun demikian, nilai *ROC/AUC C4.5* sebesar 0,835 menunjukkan bahwa meskipun akurasi klasifikasinya tinggi, kemampuan model dalam membedakan kelas positif dan negatif masih lebih rendah dibandingkan *Random Forest*. Hal ini disebabkan oleh keterbatasan pohon keputusan tunggal dalam menangani variasi dan kompleksitas data.

Sementara itu, algoritma *Naive Bayes* memperoleh nilai *accuracy* sebesar 95,59**%** dan *ROC/AUC* sebesar 0,953, yang menunjukkan performa klasifikasi yang cukup baik. Meskipun memiliki asumsi independensi antar atribut, Naive Bayes tetap mampu memberikan hasil yang kompetitif pada dataset diabetes. Algoritma ini unggul dalam kesederhanaan dan kecepatan komputasi, namun kurang optimal dalam menangkap hubungan kompleks antar variabel. Algoritma *K-Nearest Neighbor (KNN)* menunjukkan nilai *accuracy* sebesar 94,87**%** dan *ROC/AUC* sebesar

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

0,869, yang merupakan nilai terendah di antara keempat algoritma. Hal ini mengindikasikan bahwa performa *KNN* sangat dipengaruhi oleh pemilihan parameter nilai *k* serta distribusi data. Selain itu, *KNN* cenderung sensitif terhadap data yang memiliki skala dan kepadatan yang bervariasi.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa *Random Forest* merupakan algoritma yang paling direkomendasikan untuk digunakan dalam sistem prediksi faktor penyebab penyakit diabetes karena memiliki performa paling stabil dan akurat. Hasil ini juga memperkuat temuan penelitian sebelumnya yang menyatakan bahwa metode *ensemble learning* lebih efektif dibandingkan metode klasifikasi tunggal dalam menangani data medis yang kompleks.

4. KESIMPULAN

Penelitian ini melakukan pengujian model dengan membandingkan empat metode data mining, yaitu *algoritma C4.5, Naive Bayes, K-Nearest Neighbor* (*KNN*), dan *Random Forest*, menggunakan dataset yang bersumber dari komunitas global pada studi kasus *Electronic Health Records* (*EHRs*). Dataset yang digunakan berjumlah 10.000 record dengan delapan atribut prediktor dan satu atribut kelas yang menyatakan status pasien, terdiri dari 859 pasien diabetes dan 9.141 pasien non-diabetes, yang bertujuan untuk mengklasifikasikan diabetes berdasarkan riwayat kesehatan dan informasi demografis pasien. Evaluasi model dilakukan menggunakan metrik accuracy dan *Area Under Curve* (*AUC*). Hasil evaluasi menunjukkan bahwa performa masing-masing algoritma berbeda-beda bergantung pada pembagian data *training* dan data *testing*, di mana pada rasio 90%:10% dan 70%:30% algoritma C4.5 dan Random Forest menghasilkan nilai *accuracy* tertinggi dibandingkan *Naive Bayes* dan *KNN*. Selain itu, berdasarkan evaluasi *ROC curve*, *algoritma Random Forest* menunjukkan performa terbaik dengan nilai *AUC* mendekati 1, yaitu 0,972 pada pembagian data 70%:30% dan 0,970 pada pembagian data 80%:20%. Secara keseluruhan, hasil pengujian menunjukkan bahwa *algoritma C4.5* dan *Random Forest* memiliki kinerja yang hampir sama baiknya ditinjau dari nilai *accuracy* maupun *AUC*, sehingga keduanya layak digunakan dalam klasifikasi penyakit diabetes.

REFERENCES

[1] Karmila Hannum Dly, "Penerapan Data Mining Metode Algoritma C4.5 Dalam Memprediksi Tingkat Perceraian Di Kecamatan Kuranji Kota Padang Berbasis Website," *J. Sains Inform. Terap. (JSIT)*, vol. 4, no. 3, pp. 493–501, 2025, [Online]. Available: https://doi.org/10.37676/jmi.v17i1.1317

[2] R. D. Apyuma, "Penerapan Data Mining Untuk Prediksi Permintaan Hasil Pertanian Beras Menggunakan Metode FP-Growth Berbasis Website," *J. Sains Inform. Terap.*, vol. 4, no. 3, pp. 559–567, 2025, [Online]. Available: https://doi.org/10.37034/jsisfotek.v3i3.49

[3] Minyechil, "Diabetes Analysis And Prediction Using Random Forest, KNN, Naive Bayes, and J48: An Ensemble Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, pp. 348–358, 2023.

[4] I. H. Faruque, M. F., Asaduzzaman, & Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," *Int. J. Comput. Sci. Mob. Comput.*, vol. 8, no. 3, pp. 5–6, 2019.

[5] M. R. Hunafa dan A. Hermawan, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik.v4i3.1486.

[6] A., A. Divi Adiffia Freza, and J. Christina, "Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 7, no. 4, pp. 2641–2650, 2023, doi: 10.33395/sinkron.v8i4.12998

[7] S. Rizki Alifia dan H. Rahmatina, "Comparative Study of K-Nearest Neighbor and Naive Bayes for Diabetes Risk Classification," *SMATIKA : STIKI Informatika Jurnal*, vol. 14, no. 2, pp. 297–303, 2022, doi: 10.32664/smatika.v14i02.1350

[8] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Sistemasi*, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.

[9] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 122–130, 2022, [Online]. Available: http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

[10] M. R. A. A. et Al, "Perbandingan algoritma c4.5 dengan c4.5 berbasis bagging dalam menganalisa pelanggan pulsa elektronik 1," *J. Teknol. Inf. dan Multimed.*, vol. 4, no. 3, pp. 1–11, 2015, [Online]. Available: https://doi.org/10.35746/jtim.v7i4.747

[11] M. Barak, T. Ashkar, Y. J. Dori, M. Barak, T. Ashkar, and Y. J. Dori, "Teaching Science via Animated Movies : Its Effect on Students ' Learning Outcomes and Motivation," *Teh. Inform.*, vol. 1, no. 1, pp. 1–6, 2023, [Online]. Available: https://doi.org/10.56211/helloworld.v2i1.193

[12] M. A. . Andrés, "Gaming in Higher Education: Students'Assesment on Game-Based Learning. Proceedings of the 45th Conference of the International Simulation and Gaming Association," *Matematika*, vol. 3, no. 1, pp. 5–7, 2023, [Online]. Available: doi: 10.17509/cd.v9i2.11339.

[13] I. Ayu, G. Suwiprabayanti, N. Luh, P. Trisnawati, and U. Udayana, "SISTEM PENDETEKSI KESEHATAN MENTAL REMAJA MENGGUNAKAN METODE FORWARD CHAINING DAN," *J. Sist. Inf. dan Inform.*, vol. 8, no. 1, pp. 212–222, 2025, [Online]. Available: doi: 10.17509/cd.v9i2.11339.

[14] P. Ramadani, R. Fadillah, Q. Adawiyah, B. Restu, and A. Ghazali, "Perbandingan Algoritma Naïve Bayes , C4 . 5 , dan K-Nearest Neighbor untuk Klasifikasi Kelayakan Program Keluarga Harapan," *J. MEDIA Inform. [JUMIN]*, vol. 6, no. 1, pp. 775–782, 2024, [Online]. Available: https://doi.org/10.56211/helloworld.v2i1.193

[15] M. F. Nasrullah, R. R. Saedudin, and F. Hamami, "Sistem Informasi , Teknik dan Teknologi Terapan Comparison Accuracy of C4 . 5 Algorithm and K-Nearest Neighbors for Rainfall Classification," *J. SITEKNIK (Sistem Informasi, Tek. dan Teknol. Ter.*, vol. 1, no. 2, pp. 90–100, 2024.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 2118−2126
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8683

[16] M. Anshori, N. Rikatsih, M. S. Haris, "Prediksi Pasien Dengan Penyakit Kardiovaskular Menggunakan Random Forest,", *Jurnal TEKTRIKA*, vol. 7, no. 2, pp. 58–64, 2023, doi: https://doi.org/10.25124/tektrika.v7i2.5279

[17] H. G. et Al, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid," *J. Ris. Rumpun Ilmu Teh.*, vol. 1, no. 1, pp. 39–40, 2022, [Online]. Available: https://doi.org/10.55606/jurritek.v1i1.127

[18] F. N. Zamzami, A. Adiwijaya, dan M. D. P "Analisis Sentimen Terhadap Review Film Menggunakan Metode Modified Balanced Random Forest dan Mutual Information," *Jurnal Media Informatika Budidarma*, vol. 8, no. 8, p. 415-421, 2020, doi: 10.30865/mib.v5i2.2835

[19] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," *Jurnal Infomedia: Teknik Informatika, multimedia & Jaringan,* vol. 9, no. 1, pp. 16-21, 2024, doi: 10.30656/prosisko.v8i1.2848

[20] M. Kholish et. al, "Perbandingan Algoritma Random Forest dan Naive Bayes dalam Memprediksi Penyakit Diabetes," *Hubisintek: Hukum Bisnis, Sains Teknologi,* vol. 5, no. 1, pp. 322–328, 2024, doi: 10.31294/ijcit.v5i1.7951

[21] R. Irfannandhy, L. B. Handoko, and N. Ariyanto, "Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes," *Edumatic : Jurnal Pendidikan Informatika,* vol. 8, no. 2, pp. 714–723, 2024, doi: 10.29408/edumatic.v8i2.27990.