JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.520-531

ANALISIS PREDIKSI PENYAKIT DIABETES MENGGUNAKAN METODE DECISION TREE, NAÏVE BAYES, K-NEAREST NEIGHBOR DAN RANDOM FOREST

Ridho Parmonangan Manurung*[1)], Afifah Syahirah[2)], Arief Wibowo[3)]
1. Magister Ilmu Komputer, Universitas Budi Luhur, Indonesia
2. Magister Ilmu Komputer, Universitas Budi Luhur, Indonesia
3. Magister Ilmu Komputer, Universitas Budi Luhur, Indonesia

Article Info
Kata Kunci: Diabetes, Decision Tree, Naive Bayes, K-NN, Random Forest

Keywords: Diabetes, Decision Tree, Naive Bayes, K-NN, Random Forest

Article history:
Received 16 May 2025
Revised 18 June 2025
Accepted 24 June 2025
Available online 1 March 2026

DOI :
https://doi.org/10.29100/jipi.v11i1.7887

* Corresponding author.
Corresponding Author
E-mail address:
2311600544@student.budiluhur.ac.id

ABSTRAK
Diabetes merupakan penyakit kronis yang ditandai dengan meningkatnya kadar gula darah yang tidak normal. Kadar gula darah yang tinggi, usia, hipertensi, sariawan terus menerus, gatal-gatal, penglihatan yang buram, berat badan berlebihan, penurunan pendengaran, kesemutan, dan faktor lainnya adalah beberapa penyebab potensial diabetes. Jika tidak ditangani dengan cepat dan tepat, diabetes dapat mengakibatkan penyakit komplikasi seperti jantung, stroke, gagal ginjal, kebutaan hingga amputasi. Data training dan data testing diabetes yang digunakan dalam penelitian ini adalah data yang berasal dari database aplikasi rekam medis dan akan diolah menggunakan empat model algoritma. Metode klasifikasi ialah bagian dari Teknik data mining untuk melakukan prediksi. Terdapat 2460 data training dan 308 data testing. Penelitian ini bertujuan untuk mengetahui model algoritma yang paling tepat untuk memprediksi diabetes. Untuk menentukan model algoritma yang paling tepat, dapat dilihat dari tingkat akurasi, presisi dan recall dari model algoritma. Hasil penelitian menunjukkan bahwa model algoritma K-Nearest Neighbor terbukti dapat melakukan prediksi penyakit diabetes dengan baik. Diperoleh nilai akurasi model algoritma K-Nearest Neighbor sebesar 83,78%, presisi sebesar 83,35% dan recall sebesar 73,02% yang diukur dengan Confusion Matrix menggunakan Rapidminer.

ABSTRACT
Diabetes is a chronic disease characterized by an abnormal increase in blood sugar levels. High blood sugar levels, age, hypertension, persistent mouth ulcers, itching, blurred vision, excessive body weight, hearing loss, tingling, and other factors are some potential causes of diabetes. If not treated quickly and properly, diabetes can lead to complications such as heart disease, stroke, kidney failure, blindness, and even amputation. The training and testing data used in this study are sourced from a medical records application database and will be processed using four algorithm models. Classification methods are a part of data mining techniques used for prediction. There are 2460 training data and 308 testing data. This study aims to determine the most appropriate algorithm model for predicting diabetes. To determine the most suitable algorithm model, accuracy, precision, and recall of each model are evaluated. The results show that the K-Nearest Neighbor algorithm model is proven to predict diabetes effectively. The K-Nearest Neighbor model achieved an accuracy of 83,78%, precision of 83,35%, and recall of 73,02%, measured using a Confusion Matrix in RapidMiner.

.

I. PENDAHULUAN

Diabetes adalah suatu kondisi kronis yang terus-menerus ditandai dengan peningkatan kadar gula darah yang tidak normal, yang dapat mengakibatkan konsekuensi kesehatan yang parah jika tidak dikelola dengan baik [1]. Glukosa merupakan sumber energi utama bagi sel-sel tubuh dan berasal dari makanan yang kita konsumsi [2]. Diabetes terjadi ketika tubuh tidak bisa menghasilkan atau menggunakan insulin dengan baik. Insulin adalah hormon yang diproduksi oleh pankreas dan berfungsi untuk mengatur kadar gula darah [3].

Orang dengan diabetes tidak dapat menggunakan insulin dengan baik, sehingga kadar gula darah menjadi tinggi. Jika tidak ditangani dengan tepat, diabetes dapat menyebabkan komplikasi serius seperti penyakit jantung, stroke, gagal ginjal, kebutaan, dan amputasi [4]. Identifikasi dan deteksi dini diabetes dapat membantu individu dalam menerapkan tindakan pencegahan dan meningkatkan hasil kesehatan mereka secara keseluruhan. Teknik pembelajaran mesin yang sedang berkembang telah menunjukkan efektivitas dalam memprediksi dan mengklasifikasikan berbagai kondisi medis, termasuk diabetes.

Studi ini menguji kinerja empat teknik pembelajaran mesin yang banyak digunakan yaitu Decision Tree, Naïve Bayes, K-Nearest Neighbor, dan Random Forest dalam konteks prediksi diabetes. Algoritme ini telah menunjukkan kemanjuran di berbagai aplikasi diagnostik medis, dan kegunaannya dalam memperkirakan diabetes menjadi perhatian akademis yang signifikan. Metode Decision Tree menggunakan model pohon keputusan untuk membuat prediksi berdasarkan fitur masukan. Dalam konteks prediksi diabetes, fitur-fitur ini mungkin mencakup usia, berat badan, tekanan darah, dan hasil tes darah relevan lainnya [5]. Selain itu, keempatnya dikenal karena implementasinya yang relatif sederhana, interpretabilitas yang tinggi, serta performa yang kompetitif dalam dataset dengan ukuran sedang. Meskipun model lain seperti Support Vector Machine (SVM) dan metode deep learning seperti neural network juga umum digunakan dalam klasifikasi medis, algoritma tersebut membutuhkan tuning parameter yang kompleks, waktu komputasi lebih tinggi, dan interpretasi yang lebih sulit. Oleh karena itu, penelitian ini berfokus pada algoritma yang lebih ringan secara komputasi namun tetap memiliki rekam jejak kinerja yang baik dalam prediksi penyakit, guna memastikan pendekatan yang efisien dan aplikatif dalam konteks penggunaan klinis atau sistem pendukung keputusan medis berbasis data rekam medis.

Algoritma Naïve Bayes adalah teknik klasifikasi probabilistik yang memanfaatkan Teorema Bayes. Hal ini bergantung pada asumsi bahwa fitur masukan saling independen, sebuah asumsi yang sering dilanggar dalam skenario dunia nyata. Terlepas dari keterbatasan ini, metode Naïve Bayes sering kali menunjukkan kinerja yang memuaskan dalam aplikasi praktis [6]. Algoritma K-Nearest Neighbors adalah teknik klasifikasi non-parametrik yang mengandalkan identifikasi dan analisis K titik data terdekat dengan masukan untuk membuat prediksi. Dalam konteks prediksi diabetes, K-NN bekerja dengan mencari K contoh data terdekat dengan informasi pasien baru, berdasarkan metrik jarak yang ditentukan, dan kemudian mengklasifikasikan pasien baru berdasarkan kelas mayoritas di antara tetangga terdekat tersebut [7]. Random Forest adalah metode pembelajaran ansambel yang memanfaatkan kekuatan prediksi kolektif dari beberapa model pohon keputusan untuk meningkatkan akurasi dan mengurangi risiko overfitting. Pendekatan ini melibatkan pembuatan banyak pohon keputusan, yang masing-masing dilatih pada subset dataset yang dipilih secara acak. Prediksi akhir kemudian ditentukan dengan menggabungkan keluaran dari seluruh pohon konstituen, biasanya melalui skema pemungutan suara mayoritas [8].

Penelitian sebelumnya telah banyak menggunakan algoritma pembelajaran mesin seperti Decision Tree, Naïve Bayes, K-NN, dan Random Forest dalam prediksi diabetes, namun sebagian besar studi masih terbatas pada pembahasan algoritma secara individu atau tanpa melakukan perbandingan menyeluruh terhadap performa masing-masing algoritma pada dataset rekam medis riil. Kesenjangan ini menciptakan kebutuhan untuk evaluasi komprehensif terhadap berbagai metode klasifikasi dalam konteks aplikasi nyata. Penelitian ini memberikan kontribusi baru dengan menguji keempat algoritma pada dataset yang diambil langsung dari aplikasi rekam medis, serta membandingkan performa berdasarkan metrik akurasi, presisi, dan recall. Hasil penelitian ini diharapkan dapat memberikan wawasan yang lebih aplikatif dalam memilih model klasifikasi yang optimal untuk diterapkan dalam sistem pendukung keputusan di layanan kesehatan, khususnya untuk prediksi dini diabetes. Dengan demikian, studi ini tidak hanya memperkuat temuan dari penelitian sebelumnya, tetapi juga menawarkan pendekatan yang lebih praktis dan kontekstual sesuai kebutuhan dunia medis.

Tujuan dari penelitian ini adalah untuk membandingkan akurasi, presisi, dan recall keempat metode tersebut dalam memprediksi diabetes berdasarkan dataset rekam medis pasien. Dataset yang digunakan dalam penelitian ini mencakup variabel-variabel seperti usia, jenis kelamin, indeks massa tubuh, tekanan darah, dan faktor relevan lainnya yang telah diidentifikasi sebagai faktor risiko potensial diabetes.

II. METODE

A. Dataset

Dataset ini diambil dari database aplikasi rekam medis, yang merupakan aplikasi yang digunakan pada sebuah rumah sakit dan klinik. Dataset ini memiliki 2460 data dan memiliki 9 atribut. Berikut merupakan deskripsi atribut dataset dapat dilihat pada Tabel I.

TABEL I

DESKRIPSI ATRIBUT DIABETES

Atribut	Deskripsi
Pregnancies	Jumlah kehamilan
Glucose	Kadar gula
BloodPresure	Tekanan Darah
SkinThickness	Ketebalan kulit
Insulin	Kadar Insulin
BMI	Indeks massa tubuh
DiabetesPedigreeFunction	Indikator riwayat diabtes
Age	Usia
Outcome	Negatif atau Positif

B. Decision Tree

Metode klasifikasi dan prediksi yang sangat efektif dan terkenal adalah Decision Tree. Dimana metode Decision Tree mengubah banyak fakta menjadi Decision Tree yang dapat memberikan aturan. Karena aturan tersebut menggunakan bahasa yang mudah dipahami, mereka mudah dipahami dan dijelaskan dengan ulang [9].

Untuk menjelajahi data, Decision Tree berguna untuk menemukan hubungan tersembunyi antara sebuah variabel target dan berbagai variabel input potensial [10]. "Decision Tree memiliki fleksibilitas (mudah digunakan) yang membuatnya sangat menarik, terutama karena menyajikan keuntungan dari proses pengubahan konsep menjadi gambar yang telah disajikan sangat mudah dimengerti di mana cabang-cabang dari pohon merangkum klasifikasi." proses analisis data dilakukan dengan menggunakan software Rapidminer. Praproses data dilakukan untuk mengubah dataset menjadi format data yang dapat dipahami oleh Rapidminer [11].

C. Naïve Bayes

Dengan asumsi penentuan nasib sendiri yang kuat, pengklasifikasi Naïve Bayes adalah pengklasifikasi probabilistik sederhana yang didasarkan pada teorema Bayesian, yang berasal dari statistik Bayesian [12]. Nama lain untuk Naïve Bayes adalah Simple Bayes dan Independence Bayes. Ketika variabel kelas diberikan, algoritma ini dapat memprediksi kemungkinan keanggotaan kelas, serupa dengan kemungkinan pada data yang diberi label kelas. Pengklasifikasi Naïve Bayes menganggap bahwa ada atau tidak adanya fitur (atribut) tertentu dari suatu kelas tidak terkait dengan ada atau tidak adanya fitur lain. Ini adalah persamaan Naïve Bayes [13].

$$p(C|F_1 \dots \dots F_n = \frac{p(C)p(F_1 \dots F_n|C)}{p(F_1 \dots F_n)} \quad (1)$$

Beberapa keuntungan menggunakan algoritma Naïve Bayes adalah sebagai berikut:

1. Model dapat diakses dengan cepat dan sangat terukur.
2. Menyeimbangkan linear dengan jumlah baris dan prediktor.
3. Algoritma Naïve Bayes adalah parallel.
4. Dapat digunakan untuk klasifikasi biner dan multiclass.

D. K-Nearest Neighbor (KNN)

K-Nearest Neighbor adalah teknik data mining dan pembelajaran mesin yang digunakan untuk regresi dan klasifikasi. Nilai terdekat dari titik data baru ditemukan menggunakan jarak euclidean atau jarak lainnya dalam ruang fitur. Kemudian, berdasarkan mayoritas kelas dari nilai tersebut, titik data baru diklasifikasikan ke salah satu kelas atau diberikan prediksi regresi [14].

E. Random Forest

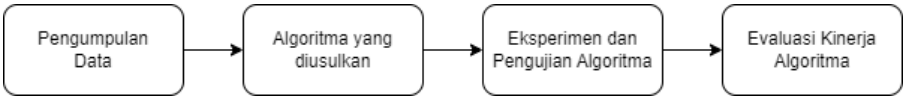
Random forest adalah salah satu metode machine learning yang digunakan untuk proses mengklasifikasikan data yang sangat besar [15]. Itu dibangun dari metode Classification and Regression Tree (CART), yang menggunakan agregasi bootstrap dan metode pemilihan fitur secara acak. Untuk mencapai keputusan akhir, Random Forest menggunakan metode kelompok untuk klasifikasi berbagai pohon melalui banyak kemunculan [16].

Random Forest dimulai dengan decision tree, teknik dasar pengolahan data. Dengan kata lain, decision tree terdiri dari sekumpulan decision tree yang digunakan untuk mengklasifikasikan data ke suatu kelas [17]. Metode ini meningkatkan kekuatan hutan dengan menumbuhkan berbagai pohon di dalamnya. Sama seperti pohon keputusan

yang menghitung entropy dan keuntungan untuk membangun satu pohon [18]. Dataset yang diambil secara acak dari data latih digunakan untuk membentuk masing-masing pohon yang akan dibentuk. Setiap pohon akan memberikan voting kelas yang paling populer selama proses klasifikasi [19].

F. Tahapan Penelitian

Seluruh tahapan yang dilakukan dalam proses penelitian ini telah dirancang secara sistematis dan divisualisasi-kan secara jelas pada Gambar 1, yang menyajikan alur atau langkah-langkah pelaksanaan penelitian dari awal hingga akhir.



Gambar. 1. Tahapan Penelitian

Pada Gambar 1, tahapan penelitian yang dilakukan sebagai berikut:

1. Pengumpulan Data
   Pengumpulan data training dan data testing berasal dari database aplikasi rekam medis. Data tersebut kemudian dilakukan preprocessing sebelum diolah menggunakan keempat algoritma.
2. Algoritma yang diusulkan
   Algoritma yang diusulkan dalam penelitian ini adalah Decision Tree, Naive Bayes, K-Nearest Neighbor dan Random Forest.
3. Eksperimen dan Pengujian Algoritma
   Eksperimen dan Pengujian Algoritma dilakukan dengan beberapa tahapan yaitu:
   a) Menyiapkan dataset untuk eksperimen.
   b) Design arsitektur untuk masing-masing algoritma yang diuji.
   c) Melakukan training terhadap setiap model dan mencatat hasil kinerja dari model.
   d) Melakukan testing terhadap setiap model dan mencatat hasil kinerja dari model
4. Evaluasi Kinerja Algoritma
   Pada saat eksperimen dan pengujian selesai, akan keluar hasil dari setiap kinerja model algoritma yang diusulkan yaitu nilai akurasi, presisi dan recall sebagai ukuran seberapa bagus model algoritma yang digunakan.

III. HASIL

A. Dataset

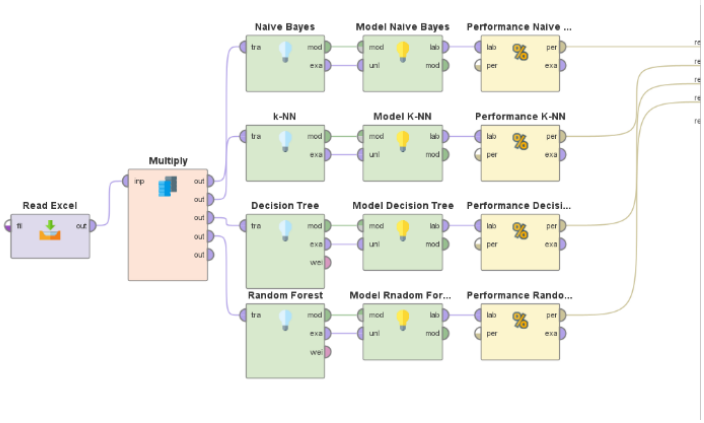
Data yang digunakan dalam penelitian berasal dari database aplikasi rekam medis. Pada Gambar 2 merupakan tangkapan layar data training diabetes yang digunakan, dari baris 1 hingga baris 2460 dan 9 atribut.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	POSITIVE
1	85	66	29	0	26.6	0.351	31	NEGATIVE
8	183	64	0	0	23.3	0.672	32	POSITIVE
1	89	66	23	94	28.1	0.167	21	NEGATIVE
0	137	40	35	168	43.1	2.288	33	POSITIVE
5	116	74	0	0	25.6	0.201	30	NEGATIVE
3	78	50	32	88	31.0	0.248	26	POSITIVE
10	115	0	0	0	35.3	0.134	29	NEGATIVE
2	197	70	45	543	30.5	0.158	53	POSITIVE
8	125	96	0	0	0.0	0.232	54	POSITIVE
4	110	92	0	0	37.6	0.191	30	NEGATIVE
10	168	74	0	0	38.0	0.537	34	POSITIVE
10	139	80	0	0	27.1	1.441	57	NEGATIVE
1	189	60	23	846	30.1	0.398	59	POSITIVE
5	166	72	19	175	25.8	0.587	51	POSITIVE
7	100	0	0	0	30.0	0.484	32	POSITIVE
0	118	84	47	230	45.8	0.551	31	POSITIVE
7	107	74	0	0	29.6	0.254	31	POSITIVE
1	103	30	38	83	43.3	0.183	33	NEGATIVE
1	115	70	30	96	34.6	0.529	32	POSITIVE
3	126	88	41	235	39.3	0.704	27	NEGATIVE
8	99	84	0	0	35.4	0.388	50	NEGATIVE
7	196	90	0	0	39.8	0.451	41	POSITIVE
9	119	80	35	0	29.0	0.263	29	POSITIVE
11	143	94	33	146	36.6	0.254	51	POSITIVE
10	125	70	26	115	31.1	0.205	41	POSITIVE
7	147	76	0	0	39.4	0.257	43	POSITIVE
1	97	66	15	140	23.2	0.487	22	NEGATIVE
13	145	82	19	110	22.2	0.245	57	NEGATIVE
5	117	92	0	0	34.1	0.337	38	NEGATIVE
5	109	75	26	0	36.0	0.546	60	NEGATIVE
3	158	76	36	245	31.6	0.851	28	POSITIVE
3	88	58	11	54	24.8	0.267	22	NEGATIVE
6	92	92	0	0	19.9	0.188	28	NEGATIVE
10	122	78	31	0	27.6	0.512	45	NEGATIVE

Gambar. 2. Data Training Diabetes

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.520-531

Pada Gambar 3 berikut merupakan tampilan untuk pengenalan data training kepada masing-masing operator diantara nya Decision Tree, Naive Bayes, KNN dan Random Forest. Digunakan juga operator Performance untuk melihat hasil akurasi dari model yang digunakan.



Gambar. 3. Data Training Algoritma

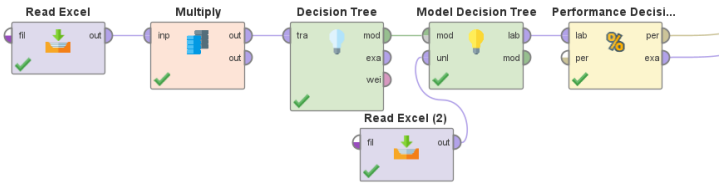
Pada Gambar 4 merupakan Data Testing yang akan diuji ke setiap model algoritma.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
9	120	72	22	56	20.8	0.733	48	
1	71	62	0	0	21.8	0.416	26	
8	74	70	40	49	35.3	0.705	39	
5	88	78	30	0	27.6	0.258	37	
10	115	98	0	0	24.0	1.022	34	
0	124	56	13	105	21.8	0.452	21	
0	74	52	10	36	27.8	0.269	22	
0	97	64	36	100	36.8	0.6	25	
8	120	0	0	0	30.0	0.183	38	
6	154	78	41	140	46.1	0.571	27	
1	144	82	40	0	41.3	0.607	28	
0	137	70	38	0	33.2	0.17	22	
0	119	66	27	0	38.8	0.259	22	
7	136	90	0	0	29.9	0.21	50	
4	114	64	0	0	28.9	0.126	24	
0	137	84	27	0	27.3	0.231	59	
2	105	80	45	191	33.7	0.711	29	
7	114	76	17	110	23.8	0.466	31	
8	126	74	38	75	25.9	0.162	39	
4	132	86	31	0	28.0	0.419	63	
3	158	70	30	328	35.5	0.344	35	
0	123	88	37	0	35.2	0.197	29	
4	85	58	22	49	27.8	0.306	28	
0	84	82	31	125	38.2	0.233	23	
0	145	0	0	0	44.2	0.63	31	
0	135	68	42	250	42.3	0.365	24	
1	139	62	41	480	40.7	0.536	21	
0	173	78	32	265	46.5	1.159	58	
4	99	72	17	0	25.6	0.294	28	
8	194	80	0	0	26.1	0.551	67	
2	83	65	28	66	36.8	0.629	24	
2	89	90	30	0	33.5	0.292	42	
4	99	68	38	0	32.8	0.145	33	

Gambar. 4. Data Testing Diabetes

B. Prediksi Decision Tree

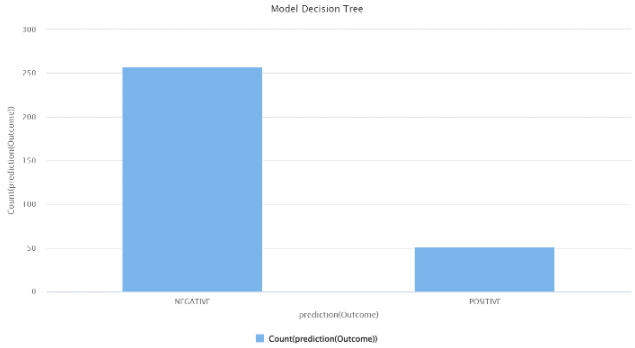
Gambar 5 merupakan pengujian data testing dengan menggunakan Decision Tree . Kemudian data testing ditunjukkan oleh Read Excel (2) dan data training ditunjukkan oleh Read Excel (1). Pengujian data testing ini dilakukan dengan menggunakan Rapidminer. Operator Performance digunakan untuk mengetahui hasil kerja atau akurasi dari model yang digunakan.



Gambar. 5. Prediksi Decision Tree Rapidminer

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.520-531

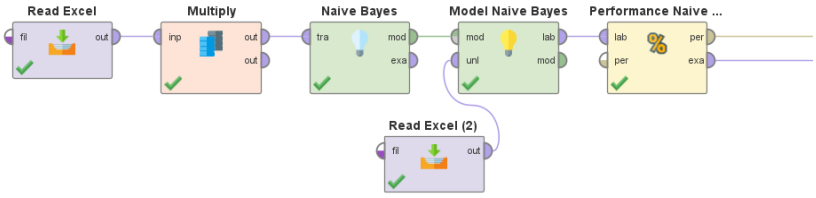
Pada Gambar 6 diagram chart berikut, dapat dilihat bahwa pada model Decision Tree memperoleh jumlah Prediksi Negative sebanyak 257. Sedangkan jumlah Prediksi Positive, diperoleh sebanyak 51 dari total data testing yaitu sebanyak 308 data.



Gambar. 6. Hasil Prediksi Decision Tree

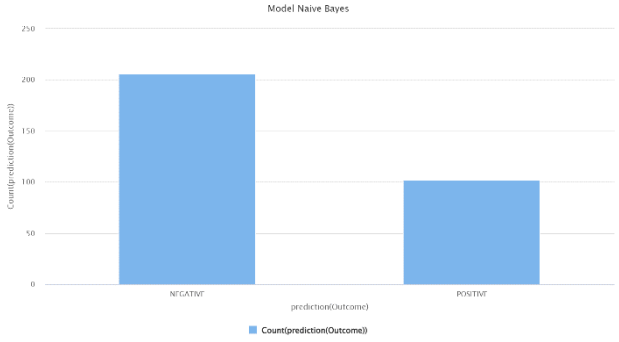
C. *Prediksi Naïve Bayes*

Kemudian pada Gambar 7 dilakukan pengujian dengan model Naive Bayes dan data testing dan data training yang sama dengan model Decision Tree. Pengujian ini juga dilakukan dengan menggunakan Rapidminer. Yang menjadi pembeda dengan yang sebelumnya adalah model yang digunakan yaitu Naive Bayes. Dan tidak lupa juga untuk ditambahkan operator Performance untuk mengukur tingkat akurasi pada model Naive Bayes [20].



Gambar. 7. Prediksi Naive Bayes Rapidminer

Pada Gambar 8 diagram chart dengan menggunakan model Naive Bayes, dapat dilihat bahwa jumlah Prediksi Negative yaitu sebanyak 206 data dan jumlah Prediksi Positive yaitu sebanyak 102 data dari total data testing yaitu 308 data.

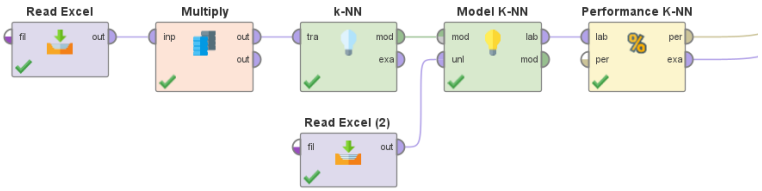


Gambar. 8. Hasil Prediksi Naive Bayes

D. *Prediksi KNN*

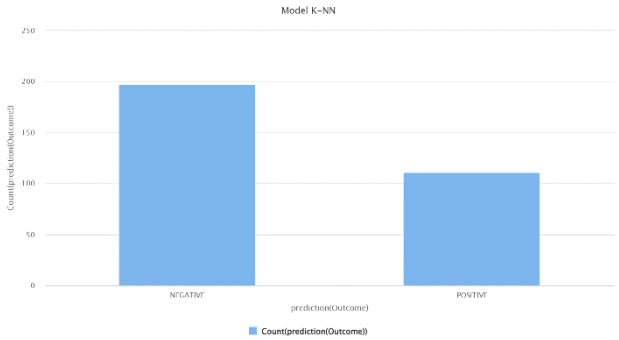
Selanjutnya pada Gambar 9 merupakan tampilan pada Rapidminer dengan menggunakan model K-Nearest Neighbor (KNN). Pengujian ini menggunakan data training yang ditandai dengan Read Excel dan data training ditandai dengan Read Excel (2). Untuk mengukur tingkat akurasi pada model KNN, dapat menggunakan operator

Performance.



Gambar. 9. Prediksi K-Nearest Neighbor Rapidminer

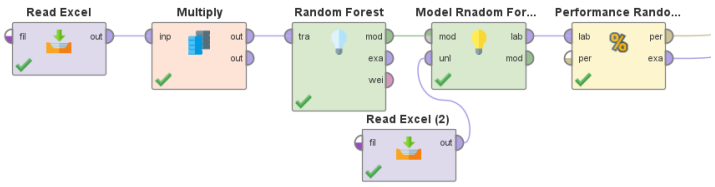
Setelah dilakukan pengujian dengan menggunakan model KNN di Rapidminer, diperoleh hasil sebagai berikut. Pada Gambar 10 diagram chart dapat dilihat bahwa Prediksi Negative diperoleh sebanyak 197 data. Sedangkan untuk Prediksi Positive sebanyak 111 data dari total data secara keseluruhan yaitu 308 data.



Gambar. 10. Hasil Prediksi K-Nearest Neighbor

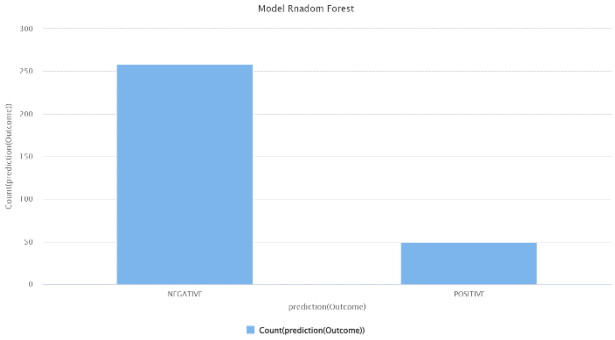
E. Prediksi Random Forest

Model pengujian yang terakhir yaitu Random Forest. Pada Gambar 11 dapat dilihat bahwa pengujian ini dilakukan di Rapidminer dan model yang digunakan adalah Random Forest. Dan tidak lupa juga untuk menambahkan operator Performance untuk mengukur tingkat akurasi dari suatu model yang akan diuji.



Gambar. 11. Prediksi Random Forest Rapidminer

Setelah model pengujian dijalankan dengan menggunakan Model Random Forest, dapat dilihat pada Gambar 12 diagram chart berikut bahwa diperoleh hasil Prediksi Negative sebanyak 259 data dan jumlah Prediksi Positive sebanyak 49 data dari total data testing yaitu sebanyak 308 data.



Gambar. 12. Hasil Prediksi Random Forest

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.520-531

IV. PEMBAHASAN

A. Hasil *Decision Tree*

Pada Tabel II berikut, dapat dilihat hasil dari confusion matrix algoritma Decion Tree dan hasil pengujian dengan data testing menggunakan aplikasi RapidMiner.

TABEL II
CONFUSION MATRIX DECISION TREE

	True Positive	True Negative	Class Precission
Pred. Positive	411	17	96,03%
Pred. Negative	549	1483	72,98%
Class Recall	42,81%	98,87%	

Keterangan:
TP (*True Positive*) = 411
TN (*True Negative*) = 1483
FP (*False Positive*) = 17
FN (*False Negative*) = 549

$$\text{Akurasi} = \frac{411 + 1483}{411 + 1483 + 17 + 549}$$
$$= 0{,}7699 = \mathbf{76{,}99\%}$$

$$\text{Recall} = TP_{rate}$$
$$= \frac{411}{411 + 549} = 0{,}4281 = \mathbf{42{,}81\%}$$

$$\textit{Specificity} = TN_{rate}$$
$$= \frac{1483}{1483 + 17} = 0{,}9887 = \mathbf{98{,}87\%}$$

$$\textit{Precision} = PPV$$
$$= \frac{411}{411 + 17} = 0{,}9603 = \mathbf{96{,}03\%}$$

$$NPV = \frac{1483}{1483 + 549}$$
$$= 0{,}7298 = \mathbf{72{,}98\%}$$

B. Hasil *Naïve Bayes*

Pada Tabel III berikut, dapat dilihat hasil dari confusion matrix algoritma Naïve Bayes dan hasil pengujian dengan data testing menggunakan aplikasi RapidMiner.

TABEL III
CONFUSION MATRIX NAÏVE BAYES

	True Positive	True Negative	Class Precission
Pred. Positive	571	265	68,30%
Pred. Negative	389	1235	76,05%
Class Recall	59,48%	82,33%	

Keterangan:
TP (*True Positive*) = 571
TN (*True Negative*) = 1235
FP (*False Positive*) = 265
FN (*False Negative*) = 389

$$\text{Akurasi} = \frac{571 + 1235}{571 + 1235 + 265 + 389}$$
$$= 0{,}7341 = \mathbf{73{,}41\%}$$

$$\text{Recall} = TP_{rate}$$
$$= \frac{571}{571 + 389} = 0{,}5948 = \mathbf{59{,}48\%}$$

$$\textit{Specificity} = TN_{rate}$$
$$= \frac{1235}{1235 + 265} = 0{,}8233 = \mathbf{82{,}33\%}$$

$$\textit{Precision} = PPV$$
$$= \frac{571}{571 + 265} = 0{,}6830 = \mathbf{68{,}30\%}$$

$$NPV = \frac{1235}{1235 + 389}$$
$$= 0{,}7605 = \mathbf{76{,}05\%}$$

C. Hasil *K-Nearest Neighbor*

Pada Tabel IV berikut, dapat dilihat hasil dari confusion matrix algoritma K-Nearest Neighbor dan hasil pengujian dengan data testing menggunakan aplikasi RapidMiner.

TABEL IV
CONFUSION MATRIX K-NEAREST NEIGHBOR

	True Positive	True Negative	Class Precission
Pred. Positive	701	140	83,35%
Pred. Negative	259	1360	84,00%
Class Recall	73,02%	90,67%	

Keterangan:
TP (*True Positive*) = 701
TN (*True Negative*) = 1360
FP (*False Positive*) = 140
FN (*False Negative*) = 259

$$\text{Akurasi} = \frac{701 + 1360}{701 + 1360 + 140 + 259}$$
$$= 0{,}8378 = \mathbf{83{,}78\%}$$

$$\text{Recall} = TP_{rate}$$
$$= \frac{701}{701 + 259} = 0{,}7302 = \mathbf{73{,}02\%}$$

$$\textit{Specificity} = TN_{rate}$$
$$= \frac{1360}{1360 + 140} = 0{,}9067 = \mathbf{90{,}67\%}$$

$$\textit{Precision} = PPV$$
$$= \frac{701}{701 + 140} = 0{,}8335 = \mathbf{83{,}35\%}$$

$$NPV = \frac{1360}{1360 + 259}$$
$$= 0{,}8400 = \mathbf{84{,}00\%}$$

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.520-531

D. Hasil *Random Forest*

Pada Tabel V berikut, dapat dilihat hasil dari confusion matrix algoritma Random Forest dan hasil pengujian dengan data testing menggunakan aplikasi RapidMiner.

TABEL V
CONFUSION MATRIX RANDOM FOREST

	True Positive	True Negative	Class Precission
Pred. Positive	444	1	99,78
Pred. Negative	516	1499	74,39%
Class Recall	46,25%	99,93%	

Keterangan:
TP (*True Positive*) = 444
TN (*True Negative*) = 1499
FP (*False Positive*) = 1
FN (*False Negative*) = 516

$$\text{Akurasi} = \frac{444 + 1499}{444 + 1499 + 1 + 516}$$
$$= 0{,}7898 = \mathbf{78{,}98\%}$$

$$\text{Recall} = TP_{rate}$$
$$= \frac{444}{444 + 516} = 0{,}4625 = \mathbf{46{,}25\%}$$

$$\textit{Specificity} = TN_{rate}$$
$$= \frac{1499}{1499 + 1} = 0{,}9993 = \mathbf{99{,}93\%}$$

$$\textit{Precision} = PPV$$
$$= \frac{444}{444 + 1} = 0{,}9978 = \mathbf{99{,}78\%}$$

$$NPV = \frac{1499}{1499 + 516}$$
$$= 0{,}7439 = \mathbf{74{,}39\%}$$

E. Perbandingan Nilai Akurasi

Berdasarkan hasil pengujian terhadap keempat algoritma, terlihat bahwa K-Nearest Neighbor (K-NN) memberikan performa terbaik dalam hal akurasi (83,78%) dan recall (73,02%). Hal ini menunjukkan bahwa K-NN lebih mampu mengidentifikasi kasus positif diabetes secara efektif dibanding algoritma lainnya. Random Forest berada di posisi kedua dalam akurasi (78,98%), namun algoritma ini memiliki presisi tertinggi (99,78%), yang berarti sangat baik dalam menghindari kesalahan klasifikasi terhadap kasus negatif.

Di sisi lain, Decision Tree menghasilkan akurasi yang cukup baik (76,99%) dan presisi tinggi (96,03%), namun recall-nya paling rendah (42,81%), menandakan bahwa banyak kasus positif tidak teridentifikasi dengan baik. Naïve Bayes memiliki kinerja paling rendah secara keseluruhan, dengan akurasi 73,41%, presisi 68,30%, dan recall 59,48%, kemungkinan disebabkan oleh asumsi independensi antar fitur yang tidak terpenuhi pada data medis.

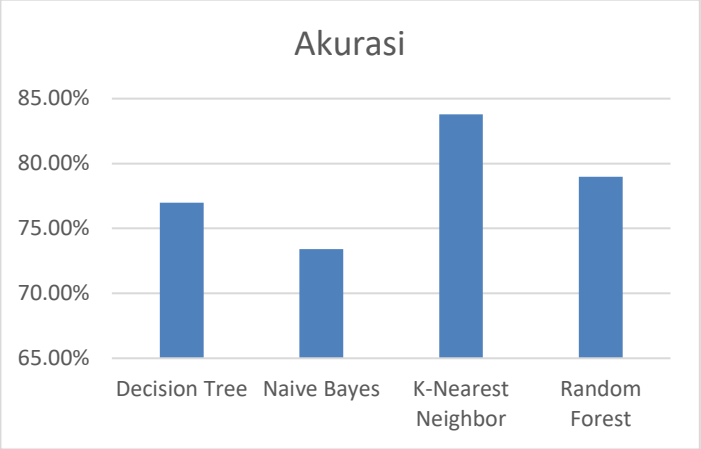
Perbandingan ini menunjukkan bahwa tidak ada satu algoritma yang unggul dalam semua metrik evaluasi. Pemilihan algoritma harus mempertimbangkan kebutuhan sistem, apakah lebih mengutamakan deteksi kasus positif (recall) seperti pada sistem skrining awal, atau menghindari kesalahan positif (presisi) seperti pada sistem diagnosis lanjutan. K-NN cocok untuk aplikasi yang membutuhkan keseimbangan antara akurasi dan kemampuan mendeteksi kasus, sedangkan Random Forest ideal untuk sistem yang membutuhkan presisi tinggi.

Untuk hasil perhitungan nilai akurasi terhadap semua model algoritma, dapat dilihat perbandingannya pada tabel VI berikut.

TABEL VI
PERBANDINGAN NILAI AKURASI

Model	Akurasi
Decision Tree	76,99%
Naïve Bayes	73,41%
K-Nearest Neighbor	83,78%
Random Forest	78,98%

Jika menggunakan grafik pada Gambar 13, dapat dilihat hasil perbandingan nilai akurasi sebagai berikut :



Gambar. 13. Grafik Perbandingan Nilai Akurasi

Peningkatan akurasi yang diperoleh dari algoritma K-Nearest Neighbor (K-NN) dibandingkan dengan tiga algoritma lainnya dapat dijelaskan melalui pendekatan klasifikasinya yang berbasis instance. Algoritma ini mengklasifikasikan data baru berdasarkan kedekatannya dengan data latih tanpa membuat asumsi statistik terhadap distribusi data. Dalam konteks dataset yang digunakan, yang memiliki atribut numerik seperti kadar glukosa, tekanan darah, dan indeks massa tubuh, pendekatan berbasis jarak seperti K-NN sangat efektif karena dapat memanfaatkan pola-pola alami dalam ruang fitur yang konsisten.

Sebaliknya, akurasi Decision Tree dan Random Forest menunjukkan kinerja yang relatif tinggi, tetapi dengan pola prediksi yang cenderung lebih konservatif terhadap kelas mayoritas. Decision Tree bekerja dengan segmentasi data berdasarkan ambang batas nilai fitur, yang dapat menyebabkan penurunan performa apabila data bersifat tidak linear atau memiliki banyak overlap antar kelas. Random Forest sedikit mengatasi keterbatasan ini dengan membangun banyak pohon, tetapi kompleksitas model yang meningkat tidak selalu linear terhadap peningkatan akurasi.

Naïve Bayes memiliki akurasi paling rendah, yang konsisten dengan asumsi dasarnya bahwa setiap fitur bersifat independen. Pada data medis, fitur-fitur seperti kadar glukosa, BMI, dan tekanan darah seringkali memiliki korelasi yang tinggi, yang menyebabkan pelanggaran terhadap asumsi independensi dan berdampak pada penurunan performa prediksi.

Dengan demikian, pola peningkatan akurasi dalam penelitian ini menunjukkan bahwa model yang mampu menangani hubungan non-linier dan keterkaitan antar fitur tanpa mengandalkan asumsi yang terlalu kuat cenderung memberikan hasil klasifikasi yang lebih baik.

Meskipun penelitian ini menunjukkan bahwa algoritma K-Nearest Neighbor memberikan hasil terbaik dalam hal akurasi dan recall, terdapat beberapa keterbatasan yang perlu diperhatikan. Pertama, dataset yang digunakan tidak dijelaskan secara rinci apakah terdapat ketidakseimbangan kelas (class imbalance), yang berpotensi mempengaruhi kinerja model, terutama pada metrik recall dan precision. Kedua, kemungkinan adanya bias pada data (misalnya bias seleksi atau representasi) tidak dibahas, yang dapat mempengaruhi generalisasi hasil penelitian. Ketiga, studi ini terbatas pada penggunaan algoritma dasar yang relatif ringan dari sisi komputasi. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi metode ensemble yang lebih kompleks seperti Gradient Boosting atau teknik deep learning seperti Artificial Neural Network (ANN) dan Convolutional Neural Network (CNN), guna meningkatkan performa prediktif. Selain itu, analisis lebih lanjut terhadap fitur yang paling berpengaruh dan penggunaan teknik seleksi fitur juga dapat memberikan wawasan yang lebih mendalam.

Hasil penelitian ini menunjukkan bahwa algoritma K-Nearest Neighbor (K-NN) memiliki akurasi tertinggi dalam memprediksi diabetes, yaitu sebesar 83,78%. Temuan ini sejalan dengan penelitian yang dilakukan oleh Anggrawan dan Mayadi (2022), yang juga menunjukkan bahwa K-NN memberikan performa yang kompetitif dalam diagnosis diabetes berbasis data rekam medis dengan akurasi di atas 80% [2]. Sebaliknya, hasil ini berbeda dengan studi yang dilakukan oleh Permana et al. (2021), yang menemukan bahwa algoritma Decision Tree lebih unggul dalam hal interpretabilitas dan akurasi pada dataset serupa [5]. Perbedaan ini kemungkinan disebabkan oleh karakteristik dataset yang digunakan, terutama dalam hal jumlah fitur dan distribusi kelas. Sementara itu, akurasi Naïve Bayes yang rendah dalam studi ini juga sejalan dengan penelitian Berrar (2019) yang menjelaskan bahwa asumsi independensi antar fitur seringkali tidak terpenuhi dalam data medis, sehingga dapat menurunkan performa model [11]. Hal ini menunjukkan bahwa pemilihan algoritma harus disesuaikan dengan karakteristik data dan kebutuhan klinis. Dengan demikian, penelitian ini memberikan kontribusi dalam memperkuat bukti

empiris bahwa algoritma K-NN dapat menjadi pilihan efektif untuk implementasi sistem pendukung keputusan klinis berbasis data riil.

Penerapan model prediksi dalam praktik klinis memerlukan sistem pendukung keputusan medis yang mampu mengolah data secara real-time dan memberikan rekomendasi yang dapat diandalkan oleh tenaga kesehatan. Salah satu tantangan utama dalam integrasi ini adalah memastikan interoperabilitas dengan sistem rekam medis elektronik yang telah ada, serta kesiapan infrastruktur teknologi di institusi layanan kesehatan. Di samping itu, isu privasi dan keamanan data pasien menjadi perhatian krusial, mengingat penggunaan data medis harus mematuhi regulasi perlindungan data seperti UU Perlindungan Data Pribadi. Model yang dikembangkan juga perlu dievaluasi secara berkelanjutan untuk menghindari bias algoritma yang dapat muncul dari data pelatihan yang tidak representatif. Oleh karena itu, keberhasilan implementasi model tidak hanya bergantung pada akurasi teknis, tetapi juga pada kelayakan operasional, kepatuhan etika, dan kesiapan ekosistem digital di dunia nyata.

V. KESIMPULAN

Dari hasil dan pembahasan penelitian, dapat ditarik kesimpulan bahwa algoritma K-Nearest Neighbor (KNN) mampu memprediksi penyakit Diabetes dengan tingkat akurasi 83,78%. Pada urutan kedua yang memperoleh tingkat akurasi tinggi yaitu algoritma Random Forest dengan 78,98%. Kemudian pada urutan ketiga yaitu algoritma Decision Tree sebesar 76,99%. Dan yang paling terakhir dengan tingkat akurasi paling rendah yaitu algoritma Naive Bayes sebesar 73,41%. Sedangkan untuk tingkat presisi tinggi yaitu algoritma Random Forest sebesar 99,78%. Kemudian algoritma Decision Tree sebesar 96,03%. Lalu algoritma K-Nearest Neighbor (KNN) sebesar 83,35%. Dan yang terakhir yaitu Naive Bayes sebesar 68,30%. Untuk hasil recall, diperoleh algoritma K-Nearest Neighbor (KNN) memiliki hasil recall tinggi, yaitu sebesar 73,02%. Kemudian algoritma Naive Bayes sebesar 59,48%. Lalu algoritma Random Forest dengan recall 46,25%. Dan yang terakhir yaitu algoritma Decision Tree sebesar 42,81%.

Dari kesimpulan diatas, maka algoritma K-Nearest Neighbor terbukti sangat handal untuk dijadian rujukan pada pengembangan model prediksi penyakit Diabetes diantara algoritma yang lain.

DAFTAR PUSTAKA

[1] G. Mahalisa and N. Arminarahmah, "Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm".
[2] A. Anggrawan and M. Mayadi, "Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes".
[3] D. L. Tahapary, S. Wafa and D. S. Harbuwono, "Puasa Ramadan dan Diabetes Melitus: Risiko, Manfaat dan Peluang Penelitian".
[4] A. Andriyani, F. Megawaty, K. Karnelasatri, J. Trisina and S. W. N. Munthe, "EDUKASI PENYAKIT DIABETES MELITUS TIPE 2 DAN PEMERIKSAAN GULA DARAH SEWAKTU".
[5] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto and I. P. Gunawan, "Classification of diabetes disease using decision tree algorithm (C4.5)".
[6] M. R. Qisthiano, T. B. Kurniawan, E. S. Negara and M. Akbar, "Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes".
[7] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial".
[8] E. Scornet, "Trees, forests, and impurity-based variable importance".
[9] Zhao, Y., & Zhang, Y., "Comparison of decision tree methods for finding active objects." arXiv preprint arXiv:0708.4274.
[10] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.
[11] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds., Academic Press, 2019, pp. 403–412, doi: 10.1016/B978-0-12-809633-8.20473-1.
[12] V. M. Herrera, T. M. Khoshgoftaar, F. Villanustre, and B. Furht, "Random forest implementation and optimization for Big Data analytics on LexisNexis's high performance computing cluster platform," *Journal of Big Data*, vol. 6, no. 1, p. 68, 2019, doi: 10.1186/s40537-019-0224-2.
[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[14] B. Talekar and S. Agrawal, "A Detailed Review on Decision Tree and Random Forest," *Bioscience Biotechnology Research Communications*, vol. 13, no. 14, pp. 245–248, 2020.
[15] M. F. Mustapha, A. Z. B. Zulkifli, O. Kairan, N. N. S. M. Zizi, N. N. Yahya and N. M. Mohamad, "The prediction of student's academic performance using RapidMiner".
[16] R. Adawiyah, M. Muljono and W. E. Nugroho, "Algoritma Naive Bayes Untuk Memprediksi Bimbingan Konseling Siswa Sekolah Menengah Kejuruan".
[17] B. Talekar and S. Agrawal, "A Detailed Review on Decision Tree and Random Forest," *Bioscience Biotechnology Research Communications*, vol. 13, no. 14, pp. 245–248, 2020.
[18] H. Xu, G. Fan and Y. Song, "Application Analysis of the Machine Learning Fusion Model in Building a Financial Fraud Prediction Model".
[19] Y. Wang and Y. Wang, "Belief Entropy Tree and Random Forest: Learning from Data with Continuous Attributes and Evidential Labels," *Entropy*, vol. 24, no. 5, p. 605, 2022, doi: 10.3390/e24050605.
[20] A. Sutrisno and A. Rahman, "Perbandingan Algoritma Decision Tree dan Naïve Bayes dalam Menentukan Predikat Kinerja Dosen," *Jurnal Teknologi dan Manajemen Informatika*, vol. 10, no. 2, pp. 123–130, 2022.