KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Andri Setiawan[1], Zikri Hadryan Nst[2], Zuriatul Khairi[3], Rahmaddeni[4], Lusiana Efrizoni[5]

[12345,]Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia

Jln. Purwodadi indah KM.10, Sidomulyo barat, Sidomulyo bar., Kec. Tampan, kota Pekanbaru, Riau 28294
[1]Andrisetiawan20182@gmail.com ,[2]Zikrihadryan12@gmail.com , [3]Zuriatulkhairi241@gmail.com ,
[4]rahmaddeni@sar.ac.id , [5]lusiana@stmik-amik-riau.ac.id

Abstract

Diabetes mellitus, a chronic metabolic disease, causes excessive blood sugar levels because the body does not produce enough insulin or is unable to use it effectively. This condition is usually caused by organ problems. Prolonged increase in blood glucose due to diabetes can lead to kidney failure, blindness, and heart attack. Diabetes is indicated in patients who have higher blood glucose than normal. The World Health Organization (WHO) says that in 2011, 4.6 million people died from diabetes, 2.2 million from high blood glucose levels, and 1.6 million from diabetes. The number of people suffering from diabetes in 2015 was 415 million, and is expected to rise to 642 million by 2040. One of the data mining methods that helps prediction is the classification of mining data; the purpose of this research is to create and apply this classification. The research also investigates the epidemiology of diabetes, risk factors, diabetes prevention and management strategies, and strategies to reduce the burden of diabetes in the future. Classification can be done using the Random Forest algorithm. Data on diabetes is processed by the authors using the random Forest method. The model evaluation results show that the Forest Random algority can accurately classify the risk of diabetes. From a selection of variable rankings, it was found that there were 300 patients with low risk of diabetes and 20 patients with high risk of diabetic. We also found that models can normally correctly classify data with accuracy values of 98%, and AUCs are categorized as "extraordinary classification" due to the ability to distinguish positive and negative classes with 100% Area Under Curve (AUC) values.

***Keywords** : Random Forest Algorithms, Data Mining, Diabetes Disease*

Abstrak

Diabetes mellitus, penyakit metabolik kronis, menyebabkan kadar gula darah berlebihan karena tubuh tidak menghasilkan cukup insulin atau tidak mampu menggunakannya secara efektif. Kondisi ini biasanya disebabkan oleh masalah organ tubuh. Peningkatan glukosa darah yang berlarut-larut akibat diabetes dapat menyebabkan gagal ginjal, kebutaan, dan serangan jantung. Diabetes dinyatakan pada pasien yang memiliki glukosa darah yang lebih tinggi dari nilai normal. Organisasi Kesehatan Dunia (WHO) mengatakan bahwa pada tahun 2011, 4,6 juta orang meninggal karena diabetes, 2,2 juta karena kadar glukosa darah tinggi, dan 1,6 juta karena diabetes. Jumlah orang yang menderita diabetes pada tahun 2015 adalah 415 juta, dan diperkirakan akan meningkat menjadi 642 juta pada tahun 2040. Salah satu metode data mining yang membantu prediksi adalah klasifikasi data mining; tujuan penelitian ini adalah untuk membuat dan menerapkan klasifikasi ini. Penelitian ini juga menyelidiki epidemiologi diabetes, faktor risiko, strategi pencegahan dan pengelolaan diabetes, dan strategi untuk mengurangi beban penyakit diabetes di masa depan. Klasifikasi dapat dilakukan dengan algoritma *Random Forest*. Data tentang penyakit diabetes diolah oleh penulis melalui metode *Random Forest*. Hasil evaluasi model menunjukkan bahwa algoritma *Forest Random* dapat mengklasifikasikan risiko diabetes dengan akurat. Dari seleksi rangking variabel, ditemukan bahwa ada 300 pasien dengan risiko diabetes rendah dan 20 pasien dengan risiko diabetes tinggi. Kami juga menemukan bahwa model biasanya dapat mengklasifikasikan data dengan benar dengan nilai akurasi sebesar 98%, dan *AUC* dikategorikan sebagai *"Excellent Classification"* karena kemampuan untuk membedakan kelas positif dan negatif dengan nilai *Area Under Curve (AUC)* 100%.

Kata kunci : *Algoritma Random Forest, Data Mining, Penyakit Diabetes*

1. PENDAHULUAN

Penyakit metabolis yang kronis, diabetes mellitus menyebabkan gula darah berlebihan, yang sering menyebabkan komplikasi pada organ tubuh, karena tubuh pasien tidak menghasilkan insulin yang cukup atau tidak dapat memanfaatkannya dengan baik. Pasien didiagnosa diabetes ketika kadar glukosa darahnya melebihi batas normal yaitu dibawah 140md/dl[1]. Penyakit diabetes disebabkan oleh peningkatan kadar glukosa dalam darah. Peningkatan terus-menerus kadar glukosa darah dapat menyebabkan komplikasi seperti gagal ginjal, kebutaan, dan serangan jantung[2].

Organisasi Kesehatan Dunia (WHO) melaporkan bahwa 4,6 juta kematian disebabkan oleh diabetes pada tahun 2011. Glukosa darah tinggi menyebabkan 2,2 juta kematian pada tahun 2012, dan diabetes menyebabkan 1,6 juta kematian pada tahun 2016. Menurut International Diabetes Federation, sebanyak 415 juta orang mengidap diabetes pada tahun 2015. Pada tahun 2040, jumlah ini diperkirakan akan meningkat sebanyak 227 juta orang, atau menjadi 642 juta orang. Biaya kesehatan terkait Diabetes Mellitus juga mencapai 465 miliar dolar [3]. Banyak orang mencatat diabetes untuk mencegah dan mengevaluasi pasien yang didiagnosis sejak dini. Salah satu cara untuk mencatatkan adalah dengan menggunakan teknik klasifikasi data mining[4]. Proses mengekstraksi informasi penting dari basis data yang sangat besar dikenal sebagai data mining. Tujuan dari proses ini adalah untuk mengubah informasi ini menjadi informasi baru yang dapat membantu pengambilan keputusan[5]. Sedangkan menurut peneliti[4], Data mining adalah proses yang menggabungkan deskripsi atau gambar, prediksi atau ramalan, clustering, klasifikasi, dan asosiasi, dan estimasi. Ini adalah teknik yang umum digunakan untuk membuat hubungan yang memiliki arti, pola, dan kecondongan dengan menggunakan teknik identifikasi pola untuk kelompok data yang sangat besar yang disimpan.

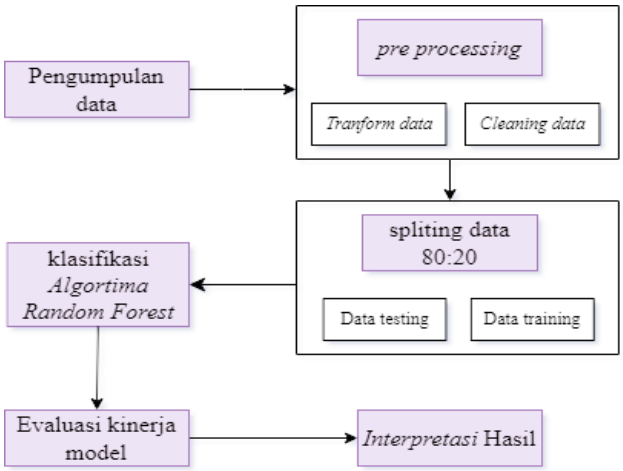
Klasifikasi adalah proses mencari himpunan fungsi (model) yang dapat menjelaskan dan membedakan kelas data atau konsep. Tujuan klasifikasi adalah untuk menggunakan himpunan model untuk memprediksi kelas objek yang belum diketahui[6]. Penelitian ini membutuhkan metode data mining dengan algoritma Random Forest untuk memulai klasifikasi.

Random Forest adalah evolusi dari metode pohon keputusan yang terdiri dari beberapa pohon keputusan; setiap pohon keputusan telah dilatih dengan sampel unik, dan setiap atribut digabungkan menjadi pohon yang dipilih dari subset atribut yang acak. *Random Forest* memiliki sejumlah manfaat. Selain menjadi tempat penyimpanan data yang efisien, ini dapat meningkatkan akurasi hasil dalam kasus data hilang dan menolak *outlier*. Selain itu, *Random Forest* memiliki proses seleksi fitur yang memungkinkan untuk memilih fitur terbaik untuk meningkatkan kinerja model klasifikasi[7].
Tingkat akurasi algoritma *Decision Tree* adalah 73,30% dengan nilai *AUC kurva ROC* 0.733, sedangkan algoritma *Naive Bayes* memiliki 75,13% dengan nilai *AUC kurva ROC* 0,810. Dengan demikian, algoritma *Naive Bayes* menunjukkan hasil prediksi yang lebih baik dalam memprediksi tingkat risiko terkena diabetes, menurut penelitian terdahulu yang telah meneliti hubungannya dengan memprediksi tingkat risiko terkena diabetes[2].

Dengan demikian, diharapkan bahwa penelitian ini dapat membantu dalam pencegahan dan pengelolaan diabetes. Diharapkan algoritma *Random Forest* dapat menghasilkan nilai deteksi penyakit diabetes yang lebih akurat dengan menggunakan dataset yang sama dengan penelitian sebelumnya dataset yang diperoleh dari situs web Kaggle. Selain itu, diharapkan bahwa penelitian ini akan berfungsi sebagai dasar untuk studi lanjutan dalam ilmu data dan teknologi kesehatan.

2. METODOLOGI PENELITIAN

Gambar 1 menunjukkan jalan penelitian: teori, hipotesis, analisis, dan rancangan.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Pertama, penelitian ini mengumpulkan data dari situs data publik Kaggle, di mana 520 record tentang pasien diabetes dikumpulkan.

2.2 *Pre-processing*

preprocessing adalah tahap Pembersihan data di mana data yang hilang dan duplikat diperiksa, dan proses pemilihan data sehingga dapat digunakan dengan cara yang lebih terstruktur dengan menghilangkan atribut yang tidak perlu, membuat data lebih sistematis, dan menghilangkan suara. Data duplikat dapat dihapus atau mean dan median dapat digunakan untuk mengimput data yang hilang[8]. Langkah pertama yaitu,

a. *Transformasi Data*

Sebelum proses klasifikasi dimulai, data kategori diubah menjadi data numerik dalam proses transformasi data.

b. *Cleaning Data*

Proses membersihkan data merupakan langkah berikutnya dalam penelitian ini. kumpulan data yang nilainya telah dibuang oleh penulis. Data harus dianalisis untuk menemukan informasi yang tidak relevan. Setiap variabel yang ditemukan di setiap baris dalam dataset memiliki nilai panjang; tidak ada nilai yang hilang.

2.3 Spliting Data

Dataset yang telah diproses dan dibersihkan sebelumnya akan dibagi menjadi dua bagian pada tahap pembagian data. Data pelatihan akan digunakan untuk menentukan nilai kekuatan model, dan data tes akan digunakan untuk mengevaluasi hasil dari nilai kekuatan model[9]. Menurut peneliti [6], Salah satu faktor yang memengaruhi seberapa baik model klasifikasi berfungsi pada algoritma pembelajaran mesin adalah model splitting data yang digunakan untuk mempartisi dataset. Untuk memastikan bahwa model yang dibangun dapat diterapkan pada data yang baru diterima, berbagi data adalah tujuan. Kami memilih pembagian data terbesar, 80:20, dalam proses penelitian ini. Data akan diuji dengan *Random Forest.*

2.4 Klasifikasi

Klasifikasi adalah proses menemukan pola bertujuan untuk memperkirakan kelas objek yang belum diketahui untuk membuktikan bahwa sebuah objek data termasuk dalam jenis yang telah dideskripsikan sebelumnya[10]. Menurut peneliti [11], Klasifikasi adalah jenis analisis data dalam data mining yang memungkinkan prediksi berdasarkan label kelas sampel yang diklasifikasikan. Nilai-nilai atribut prediktor dipengaruhi oleh kemungkinan nilai atribut target, yang menentukan kelas yang diprediksi. Berbagai kemungkinan nilai tersedia untuk setiap instan data[12].

2.5 Algoritma *Random Forest*

Random Forest (RF) adalah pendekatan yang didasarkan pada konsep pembelajaran kelompok, di mana berbagai pohon keputusan dibangun dan digabungkan untuk menghasilkan prediksi akhir. [13]. Menurut [14], Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node*, dan *leaf node* dengan menggunakan data dan atribut secara acak sesuai dengan persyaratan; simpul anak dibangun secara acak untuk setiap *node*, meningkatkan akurasi hasil. Pembentukan algoritma ini ditentukan oleh nilai vektor acak yang independen dan seragam di setiap pohon. Metode klasifikasi adalah yang paling akurat saat membuat prediksi. Ini dapat menemukan perbandingan antar hutan acak seperti properti metode ansambel dan dapat menangani banyak variabel input tanpa *overfitting.*

Menurut [15] bahwa random forest memiliki dua konsep yaitu :

a. Membangun ensemble.
   Ensemble mencari solusi prediksi dengan hasil terbaik.

b. Penyeleksian fitur.
   Setiap pohon yang telah dibangun memiliki fitur yang dipilih secara acak.

Salah satu keuntungan menggunakan algoritma *Random Forest* sebagai metode klasifikasi adalah bahwa itu tidak menimbulkan masalah overfitting dalam classifier. *Random Forest* biasanya digunakan untuk menentukan fitur utama yang akan digunakan dari kumpulan data pelatihan dan regresi[16].

2.6 Evaluasi Kinerja Model

Dalam pembelajaran mesin, evaluasi model digunakan untuk menilai kinerja model klasifikasi. Sangat penting untuk melakukan evaluasi model untuk mengetahui kemampuan mereka untuk membuat klasifikasi yang akurat. Dalam penelitian ini, lima komponen terdiri dari laporan metrix evaluasi: *akurasi, ketepatan, recall, skor F1*, dukungan, dan konsistensi matriks. Tabel Matriks Konflik memungkinkan penulis untuk melihat kinerja algoritma pembelajaran yang diawasi. Meskipun setiap baris mengandung contoh kelas sebenarnya, setiap kolom matiks menampilkan instance dalam kelas yang diharapkan[17].

2.6.1 *Confunssion Matrix*

Konfunssion Matrix adalah alat mudah dan efektif untuk menunjukkan kinerja pengklasifikasi. Hasilnya mudah dipahami, yang merupakan keuntungan. *Matriks konfusi* dapat digunakan untuk mengevaluasi kinerja model atau algoritma [18]. Menurut peneliti [4], hasil klasifikasi diwakili oleh empat kata dalam pengukuran kinerja menggunakan *matrix confusion*. Keempat istilah tersebut adalah sebagai berikut:

1. Data Positif Palsu (FP), yang dianggap negatif tetapi diprediksi positif;
2. Data Negatif Palsu (FN), yang dianggap negatif tetapi diprediksi positif;
3. Data Positif Benar (TP), yang diprediksi benar;
4. Data Negatif Benar (TN), yang diprediksi dengan benar.

TABLE I. *CONFUSSION MATRIX*

Classification	*Predicted class*		
		Class : yes	*Class : No*
Observed Class	*Class Yes*	*A(TP)*	*B(FN)*
	Class No	*C(FP)*	*D(TN)*

Untuk menghitung akurasi digunakan rumus sebagai berikut:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \text{.....(1)}$$

Jumlah prediksi yang tepat (TP + TN) dibagi dengan jumlah total sampel untuk mendapatkan akurasi, yang ditunjukkan oleh (1). (2) dan (3) menunjukkan metode untuk merangkum hasil dalam matriks kebingungan yang paling umum.

$$Precision = \frac{TP}{TP+FP} \text{.....(2)}$$

Presisi adalah indikator kinerja ketika tujuannya adalah untuk mengurangi jumlah positif palsu (FP). Nilai prediksi positif dihitung sebagai jumlah sampel yang benar positif (TP + FP) yang diprediksi positif.

$$Sensitivity = \frac{TP}{TP+FN} \text{.....(3)}$$

Sensitivitas ditentukan dengan menghitung berapa banyak sampel positif (TP + FN) yang dianggap kelas positif (TP).

2.6.2 *ROC Curve*

Kurva ROC digunakan untuk menentukan parameter model yang diinginkan berdasarkan karakteristik model classifier. Dengan demikian, metode klasifikasi dapat dievaluasi dengan standar seperti interpretabilitas, skabilitas, akurasi, kecepatan, dan kehandalan[19]. Peneliti [20], menyatakan bahwa *ROC Curve* adalah alat visual yang membantu untuk membandingkan dua model klasifikasi. Kurva Nilai *ROC (Receiver Operating Characteritics)* umumnya digunakan untuk menilai hasil prediksi dalam bentuk grafik. Ini adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka. Untuk klasifikasi data mining, nilai *AUC* dapat dibagi menjadi berbagai kelompok[2].

a) 0.90-1.00 = *Excellent Classification*
b) 0.80-0.90 = *Good Classification*
c) 0.70-0.80 = *Fair Classification*
d) 0.60-0.70 = *Poor Classification*
e) 0.50-0.60 = *Failur*

Area di bawah kurva *(AUC)* dihitung untuk mengukur perbedaan dalam kinerja metode yang digunakan. Rumus untuk menghitung *AUC* adalah sebagai berikut.

$$\theta^y = \frac{1}{mn}\sum_i^m = 1\psi(xi^y, xj^y \text{..... (4)}$$

Dimana

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases} \quad \text{..... (5)}$$

Keterangan :

X = Output *Positif*
Y = Output *Negatif*

3. HASIL DAN PEMBAHASAN

Implementasi algoritma *Random Forest* untuk mengklasifikasikan tingkat risiko penyakit diabetes dibahas dalam bagian ini. Model dibangun, diterapkan, dan diuji dengan membagi data menjadi set pelatihan dan pengujian. dievaluasi dengan menggunakan *Matrix Confusin* dan *ROC AUC.*

3.1 Pengumpulan Data

Proses berikut menggunakan dataset yang diperoleh dari https://www.kaggle.com/datasets/rcratos/diabetes-risk-prediction/data, yang berisi 520 data dengan 16 variabel dan 1 label, seperti yang ditunjukkan pada Gambar 2.

Keterangan :

a. *Age*: Ini mewakili usia individu dalam tahun.
b. *Gender*: Ini adalah jenis kelamin individu. Bisa jadi laki-laki atau perempuan.
c. *Polyuria*: Ini mengacu pada adanya buang air kecil yang berlebihan, yang merupakan gejala umum diabetes.
d. *Polydipsia*: Ini mengacu pada haus yang berlebihan, gejala umum lainnya dari diabetes.
e. *Sudden weight loss*: Ini menunjukkan apakah individu telah mengalami penurunan berat badan yang tidak dapat dijelaskan, yang dapat menjadi tanda diabetes.
f. *Weakness*: Ini menunjukkan apakah individu mengalami kelemahan fisik umum, gejala potensial diabetes.
g. *Polyphagia*: Ini mengacu pada kelaparan yang berlebihan, gejala potensial lain dari diabetes.
h. *Genital thrush*: Ini adalah infeksi ragi yang dapat menyebabkan gatal, rasa sakit, dan ketidaknyamanan lainnya di daerah genital. Ini bisa lebih umum pada orang dengan diabetes.
i. *Visual blurring*: Ini menunjukkan apakah individu mengalami penglihatan yang kabur, gejala potensial diabetes.
j. *itching*: Ini menunjukkan apakah individu mengalami gatal umum, yang dapat menjadi gejala diabetes.
k. *Irritability*: Ini menunjukkan apakah individu mengalami iritabilitas, yang dapat menjadi gejala diabetes.
l. *Delayed healing*: Ini menunjukkan apakah individu mengalami penyembuhan luka yang lambat, yang dapat menjadi gejala diabetes.
m. *Partial paresis*: Ini mengacu pada hilangnya parsial gerakan sukarela, yang dapat menjadi gejala diabetes.
n. *Muscle stiffness*: Ini menunjukkan apakah individu mengalami kekakuan otot, yang dapat menjadi gejala diabetes.
o. *Alopecia*: Ini mengacu pada rambut rontok, yang dapat menjadi gejala diabetes.
p. *Obesity*: Ini menunjukkan apakah individu itu obesitas, yang merupakan faktor risiko utama untuk diabetes.

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	Female	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Positive
516	48	Female	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Positive
517	58	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Positive
518	32	Female	No	No	No	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	No	Negative
519	42	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative

520 rows × 17 columns

Gambar 2. Dataset

3.2 *Pre-processing*

Preprocessing membantu merangkai langkah-langkah yang diambil untuk mengubah data mentah menjadi data yang relevan untuk digunakan dalam modeling.

dtype= object)

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1
516	48	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1
517	58	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	1
518	32	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0
519	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

520 rows × 17 columns

Gambar 3. Tranformasi Data

a. Transformasi Data

Dalam Gambar 3, Tahap ini mengubah data teks menjadi data numerik untuk mempermudah dan meningkatkan kualitas data untuk analisis yang lebih akurat.

b. Cleaning Data

Setelah melakukan proses perbaikan data pada 520 daftar dengan 16 atribut, kami memastikan bahwa tidak ada nilai yang hilang. Semua data telah diverifikasi dan siap untuk digunakan untuk analisis lebih lanjut, jadi langkah berikutnya adalah klasifikasi.

3.3 Klasifikasi

Setelah melakukan tahapan preprocessing data melalui transformasi dan pembersihan data, kita kemudian melanjutkan ke tahapan klasifikasi dengan menggunakan pembagian data 80:20 pada algoritma Random Forest.

3.3.1 *Random Forest*

Fungsi *Random Forest Classifier* dari *library sklearn.ensemble* digunakan untuk melatih model, yang membantu pemodelan algoritma untuk membuat hasil klasifikasi lebih akurat.

```
[17] # Membuat model Random Forest
     rf_model = RandomForestClassifier(random_state=45)

     # Melatih model menggunakan data latih
     rf_model.fit(X_train, y_train)

     RandomForestClassifier
     RandomForestClassifier(random_state=45)

[19] # Melakukan prediksi menggunakan data uji
     rf_predictions = rf_model.predict(X_test)

     y_pred=rf_model.predict(X_test)
     y_pred

     array([0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1,
            1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1,
            0, 1, 0, 1, 1, 1, 0, 0])
```

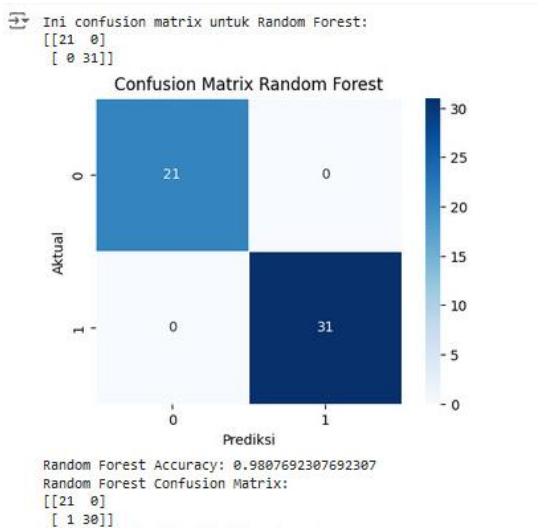
Gambar 4. *Script Random Forest*

Model Hutan Random memulai tugas klasifikasi dengan kode *rf_model = RandomForestClassifier (random_state=4 5)* dan mengatur biji untuk memastikan proses pengacakan model dapat direplikasi, seperti yang ditunjukkan pada Gambar 4. Ini adalah langkah pertama dalam proses pembuatan model. Setelah itu, model akan dilatih dan dievaluasi untuk melakukan prediksi dengan data yang diberikan.

Mengubah *rf_model.Fit(X_train, y_train)* adalah komponen penting dari proses pembelajaran mesin di mana model *Random Forest* dilatih pada data pelatihan. Metode fit memungkinkan model untuk mempelajari hubungan antara fitur dan target dalam data pelatihan, yang memungkinkan model untuk membuat prediksi yang akurat tentang data yang akan datang.

3.4 Evaluasi Kinerja Model

Hasil evaluasi untuk model klasifikasi algoritma *Random Forest* menggunakan *Confusion Matrix* dan *kurva ROC berikut* adalah hasil dari pengujian di atas:



Gambar 5. C*onfussion Matrix*

Berdasarkan Table 2 dan Gambar 5 mengguanakan metode *Random Forest* mendapatkan nilai *accuracy* 98% pada spliting data 80:20. dapat di hitung nilai *accuracy* sebagai berikut :

TP = 30 FP = 0
TN = 21 FN = 1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{30 + 21}{30 + 21 + 0 + 1}$$
$$= \frac{51}{52}$$
$$= 98\%$$

Berdasarkan hasil perhitungan, tingkat akurasi dari algortima *Random Forest* sebesar 98%.

3.4.2 ROC Curve

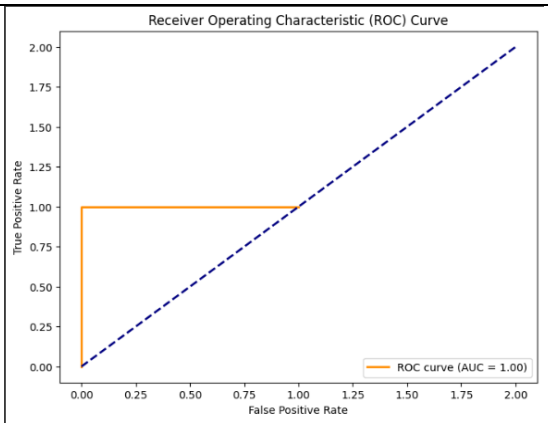
Pada Gambar 9 merupakan hasil Grafik *ROC* dengan algoritma *Random Forest.*

3.4.1 C*onfussion Matrix*

Tabel 2 dan Gambar 5 di bawah menunjukkan hasil *Matriks Konfusio* dari algoritma yang menggunakan *Random Forest.*

TABLE II. ACCURACY *RANDOM FOREST*

	precision	*Recall*	*F1-score*	*support*
0	0.95	1.00	0.98	21
1	1.00	0.98	0.98	31
Accuracy			0.98	52
Macro avg	0.98	0.98	0.98	52
Weighteda v	0.98	0.98	0.98	52



Gambar 6. *ROC Random Forest*

Gambar 6 menunjukkan grafik *ROC* dengan nilai *AUC (Area Under Curve)* sebesar 100% yang dihasilkan oleh algoritma *Random Forest.* Nilai *AUC*, yang telah dijelaskan oleh peneliti 2 sebelumnya, termasuk dalam kategori "*Excellent Classification"* karena memiliki nilai 100%.

3.4.3 Perbandingan *Accuracy* dan *AUC*

Pada tahap ini merupakan hasil dari perbandingan antara *Accuracy* dan *AUC* tedapat pada Tabel 3.

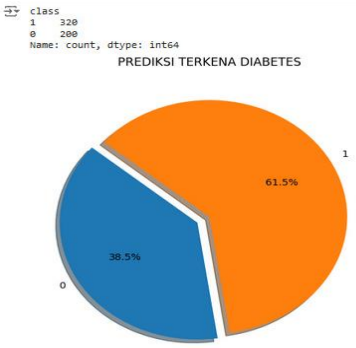
TABLE III. *HASIL ACCURACY DAN AUC*

	Algoritma Random Forest
Accuracy	*98%*
AUC	*100%*

Setelah melakukan evaluasi berbagai spliting data, penelitian ini menemukan nilai spliting data terbaik, yaitu 80:20, dengan hasil akurasi dan *AUC* metode 80:20, menurut Tabel 3. *Random Forest* mendapatkan hasil *accuracy* 98% dan *AUC* 100%.

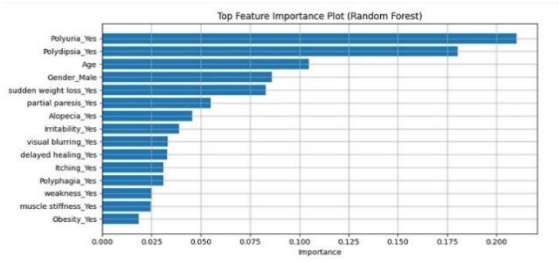
3.5 *Interprestasi* Hasil

Pada tahap ini, algoritma *Random Forest* yang digunakan untuk menganalisis hasil dari proses pengklasifikasian risiko diabetes.



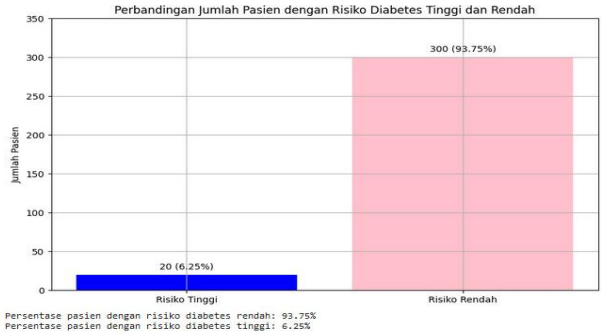
Gambar 7. Diagram Diabetes

Pada Gambar 7 diatas merupakan hasil visualisasi data yang telah dilakukan. Terlihat lebih banyak pasien menghidap *positif diabetes* sebanyak 61,5% dan *negatif diabetes* sebanyak 38,5% pasien. grafik di atas juga menyimpulkan bahwa pasien *negative diabetes* berjumlah 200 pasien, dan pasien *positif diabetes* berjumlah 320 pesien.



Gambar. *8 Rank Selection*

Gambar 8 di atas menunjukkan hasil pemilihan rangking dari data diabetes; ini menunjukkan bahwa variabel yang paling mempengaruhi pasien diabetes adalah *Polyuria, Polydipsia, Age, Gender, Sudden weight loss.*



Gambar 9. Tingkat Risiko Diabetes

Pada Gambar 9 diatas, merupakan total hasil keseluruhan prediksi terkena diabetes berjumlah 320 pasien. Dimana data telah diselection dari rangking variable yang paling mempengaruh pada data pasien diabetes yaitu *Polyuria, Polydipsia, Age, Gender, Sudden weight loss*, pada gambar diatas menjelaskan bahwa pasien yang menderita resiko diabetes rendah sebesar 300 pasien sedangkah pasien risiko diabetes tinggi sebesar 20 pasien.

4. KESIMPULAN DAN SARAN

Kami menggunakan *algoritma Random Forest* untuk mengembangkan model klasifikasi risiko diabetes dalam penelitian ini. Data yang digunakan untuk penelitian ini dibagi menjadi dua bagian, dengan perbandingan 80:20 untuk pelatihan dan pengujian model. Hasil evaluasi model menunjukkan bahwa algoritma *Random Forest* dapat dengan akurat mengklasifikasikan risiko diabetes. Kami menemukan nilai akurasi sebesar 98%, menunjukkan bahwa model biasanya mampu mengklasifikasikan data dengan benar. Selain itu, *Area Under Curve (AUC)* memiliki nilai 100% dan ditunjukkan sebagai kategori *"Excellent Classification"* karena memiliki kemampuan yang luar biasa untuk membedakan antara kelas positif dan negatif. Hasil ini dapat membantu menemukan dan mengelola diabetes pada awalnya, yang dapat meningkatkan kualitas hidup pasien dan mengurangi beban penyakit secara keseluruhan. Algoritma *Random Forest* telah ditunjukkan sebagai alat yang efektif dan dapat diandalkan untuk digunakan dalam klasifikasi risiko diabetes; data yang dikumpulkan menunjukkan bahwa ada 300 pasien dengan risiko diabetes rendah dan 20 pasien dengan risiko diabetes tinggi.

Daftar Pustaka:

[1] N. Nurussakinah and M. Faisal, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree," *J. Inform.*, vol. 10, no. 2, pp. 143–149, Oct. 2023, doi: 10.31294/inf.v10i2.15989.

[2] F. Sistem Informasi STMIK Nusa Mandiri Jakarta Jl Damai No, W. Jati Barat, and J. Selatan, "PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES," *J. Techno Nusa Mandiri*, vol. XIII, no. 1, p. 50, 2016.

[3] D. Nurul Anisa, "KLASIFIKASI PENYAKIT

DIABETES MENGGUNAKAN ALGORITMA NAIVE BAYES," *Din. Inform.*, vol. 14, no. 1, 2022.

[4] U. M. Kudus, J. Ganesha, and P. Kudus, "Fida Maisa Hana."

[5] A. K. Wahyudi, N. Azizah, and H. Saputro, "DATA MINING KLASIFIKASI KEPRIBADIAN SISWA SMP NEGERI 5 JEPARA MENGGUNAKAN METODE DECISION TREE ALGORITMA C4.5." [Online]. Available: https://journal.unisnu.ac.id/JISTER/

[6] N. Nurdiana and A. Algifari, "STUDI KOMPARASI ALGORITMA ID3 DAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS".

[7] Arifin Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, and Ricky Firmansyah, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *J. Ilm. Tek. Inform. dan Komun.*, vol. 3, no. 2, pp. 27–41, Jun. 2023, doi: 10.55606/juitik.v3i2.472.

[8] A. Setiawan, R. Febrio Waleska, M. Adji Purnama, and L. Efrizoni, "KOMPARASI ALGORITMA K-NEAREST NEIGHBOR (K-NN), SUPPORTVECTOR MACHINE (SVM), DAN DECISION TREE DALAM KLASIFIKASIPENYAKIT STROKE," 2024. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jir eISSN.2620-6900

[9] M. Asjad Adna Jihad and W. Astuti, "Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest."

[10] M. Nazar Yuniar, "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 243–246, 2023, doi: 10.55338/saintek.v5i1.1383.

[11] M. Yudhi Putra and D. Ismiyana Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI."

[12] S. Dini Widianti *et al.*, "Jurnal Informatika dan Rekayasa Perangkat Lunak Menentukan Nilai Gizi pada Balita Menggunakan Algoritma Support Vektor Machine (SVM) di Posyandu Kelurahan Ciherang".

[13] M. Salabil and N. L. Azizah, "Implementation of Data Mining in Diabetes Disease Prediction Using Random Forest and XGBoost Methods [Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost]."

[14] V. Wanika Siburian, J. Sistem Komputer Universitas Sriwijaya Palembang, and I. Elvina Mulyana, *Prediksi Harga Ponsel Menggunakan Metode Random Forest*. 2018.

[15] N. Giarsyani, "Komparasi Algoritma Machine Learning dan Deep Learning untuk Named Entity Recognition : Studi Kasus Data Kebencanaan," *Indones. J. Appl. Informatics*, vol. 4, no. 2, p. 138, 2020, doi: 10.20961/ijai.v4i2.41317.

[16] "Perbandingan Akurasi Metode Naïve Bayes Classifier dan Random Forest Menggunakan Reduksi Dimensi Linear Discriminant Analysis (LDA) untuk Diagnosisi Penyakit Diabetes".

[17] F. Pratama, Z. Hadryan Nst, Z. Khairi, and L. Efrizoni, "PERBANDINGAN ALGORITMA RANDOM FOREST DAN K-NEAREST NEIGHBOR DALAM KLASIFIKASI KESEHATAN MENTAL MAHASISWA."

[18] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2407–2413, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.

[19] S. Dewi, "KOMPARASI 5 METODE ALGORITMA KLASIFIKASI DATA MINING PADA PREDIKSI KEBERHASILAN PEMASARAN PRODUK LAYANAN PERBANKAN".

[20] L. Hermawanti, "PENERAPAN ALGORITMA KLASIFIKASI C4.5 UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA."