KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST

By Andri Setiawan

WORD COUNT 3135 TIME SUBMITTED 02-AUG-2024 08:46AM

PAPER ID 110826940

Template ini berlaku untuk terbitan Volume 6 Nomor 1 April 2023

KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Andri Setiawan

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. Organ problems usually cause this condition. Diabetes causes prolonged increases in blood glucose, which can lead to serious complications such as kidney failure, blindness, and heart attack. When patients' blood glucose levels exceed normal limits, they are to have diabetes. 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. The research is aimed at making classifications and applying data mining classification, which is one of the data-mining techniques that helps prediction. The classification can be done using Random Forest algorithms. In addition, the study investigates diabetes epidemiology, risk factors, methods of prevention and management of diabetes, and ways to reduce the burden of diabetes in the future. The author used the Random Forest method to process data on diabetes in this study. The model evaluation results show that Random Forest's algorithm can accurately classify the risk of diabetes. We found that the model can normally correctly classify the data with an accuracy of 98%. Moreover, the AUC is categorized as "Excellent classification" because it has an extraordinary ability to distinguish between positive and negative classes with an Area Under Curve (AUC) value of 100%.

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

Abstrak

Diabetes mellitus, penyakit metabolik kronis, menyebabkan kadar gula darah berlebihan karena tubuh tidak menghasilkan insulin cukup atau tidak mampu menggunakannya secara efektif. Masalah organ tubuh biasanya menyebabkan kondisi ini. Diabetes menyebabkan peningkatan glukosa darah yang berlarut-larut, yang dapat menyebabkan komplikasi serius seperti gagal ginjal, kebutaan, dan serangan jantung. Ketika glukosa darah pasien melebihi batas normal, mereka dinyatakan menderita diabetes. Organisasi Kesehatan Dunia (WHO) mengatakan bahwa pada tahun 2011, 4,6 juta orang meninggal karena diabetes, 2,2 juta orang meninggal karena kadar glukosa darah tinggi, dan 1,6 juta orang meninggal karena diabetes. Jumlah orang yang menderita diabetes pada tahun 2015 adalah 415 juta, dan diperkirakan akan meningkat menjadi 642 juta pada tahun 2040. Penelitian ini bertujuan untuk membuat klasifikasi dan menerapkan klasifikasi data mining, yang merupakan salah satu teknik data mining yang membantu prediksi. Klasifikasi dapat dilakukan dengan algoritma *Random Forest*. Selain itu, penelitian ini menyelidiki epidemiologi diabetes, faktor risiko, metode pencegahan dan pengelolaan diabetes, dan cara-cara untuk mengurangi beban penyakit diabetes di masa depan. Penulis menggunakan metode *Random Forest* untuk mengolah data tentang penyakit diabetes dalam penelitian ini. Hasil evaluasi model menunjukkan bahwa algoritma *Random Forest* dapat dengan akurat mengklasifikasikan risiko diabetes. Kami menemukan bahwa model biasanya dapat mengklasifikasikan data dengan benar dengan nilai akurasi sebesar 98%. Selain itu, *AUC* dikategorikan "*Excellent Classification*" karna memiliki kemampuan yang luar biasa untuk membedakan kelas positif dan negatif dengan nilai *Area Under Curve (AUC)* 100%. Dari seleksi rangking variabel, ditemukan bahwa terdapat 300 pasien dengan risiko diabetes rendah dan 20 pasien dengan risiko diabetes tinggi.

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

1. PENDAHULUAN

Diabetes mellitus adalah penyakit metabolis yang kronis di mana tubuh pasien tidak menghasilkan

jumlah insulin yang cukup atau tubuh pasien tidak sanggup memanfaatkan insulin dengan baik, menyebabkan gula darah berlebihan dalam tubuh, yang sering terjadi setelah komplikasi pada organ tubuh. Saat kadar glukosa darahnya melebihi batas normal, pasien didiagnosa diabetes[1]. Peningkatan kadar glukosa dalam darah menyebabkan penyakit diabetes. Peningkatan terus-menerus dalam kadar glukosa darah dapat menyebabkan komplikasi seperti gagal ginjal, kebutaan, dan serangan jantung[2].

Organisasi Kesehatan Dunia (WHO) menyatakan bahwa pada tahun 2011, diabetes menyebabkan 4,6 juta kematian; pada tahun 2012, glukosa darah tinggi menyebabkan 2,2 juta kematian, dan pada tahun 2016, diabetes menyebabkan 1,6 juta kematian. International Diabetes Federation melaporkan bahwa jumlah orang yang mengidap diabetes sebanyak 415 juta pada tahun 2015. Jumlah ini diperkirakan akan meningkat sebanyak 227 juta orang, atau menjadi 642 juta orang pada tahun 2040. Selain itu, biaya kesehatan yang terkait dengan Diabetes Mellitus telah mencapai 465 miliar dolar [3]. Pencatatan diabetes banyak dilakukan untuk mencegah dan mengevaluasi pasien yang didiagnosis sejak dini. Teknik klasifikasi data mining adalah salah satu cara untuk mencatatkan[4]. Data mining adalah proses mengekstraksi informasi berharga dari basis data yang sangat besar. Ini dilakukan untuk mengubahnya menjadi informasi baru yang dapat membantu pengambilan keputusan[5]. Sedangkan menurut peneliti[4], Data mining adalah suatu proses yang menggabungkan deskripsi atau gambar, prediksi atau ramalan, clustering, klasifikasi dan asosiasi, dan estimasi. Ini adalah prosedur yang biasa digunakan untuk menghasilkan ikatan yang memiliki arti, pola, dan kecondongan melalui penggunaan teknik identifikasi pola untuk kelompok data besar yang disimpan.

Klasifikasi adalah proses mencari suatu himpunan fungsi (model) yang dapat menjelaskan dan membedakan kelas-kelas data atau konsep-konsep. Tujuan dari klasifikasi adalah untuk menggunakan himpunan model ini untuk memprediksi kelas objek yang belum diketahui[6]. Pada penelitian ini, metode data mining yang menggunakan algoritma *Random Forest* diperlukan untuk memulai klasifikasi.

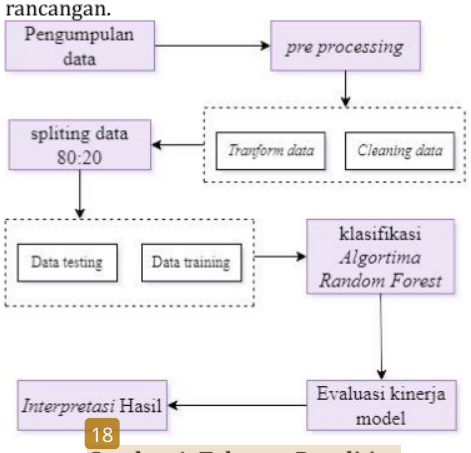
Random Forest adalah evolusi dari teknik pohon keputusan yang terdiri dari beberapa pohon keputusan; setiap pohon keputusan telah dilatih menggunakan sampel individual, dan setiap atribut digabungkan menjadi pohon yang dipilih dari subset acak atribut. *Random Forest* memiliki banyak keunggulan. Ini dapat meningkatkan akurasi hasil jika data hilang dan menolak outlier, serta menjadi tempat penyimpanan data yang efisien. Selain itu, *Random Forest* memiliki proses seleksi fitur yang memungkinkan untuk memilih fitur yang paling efektif untuk meningkatkan kinerja model klasifikasi[7].

Menurut penelitian terdahulu yang telah meneliti hubungannya dengan memprediksi tingkat risiko terkena diabetes, peneliti [2] merangkum hasil klasifikasi data evaluasi dengan menggunakan *Confusion Matrix* dan *kurva ROC*. 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. Oleh karena itu, algoritma *Naive Bayes* memiliki hasil prediksi yang baik dalam memprediksi penyakit diabetes seorang pasien.

Berdasarkan hal-hal yang disebutkan di atas, diharapkan bahwa penelitian ini dapat memberikan kontribusi yang signifikan dalam upaya pencegahan dan pengelolaan diabetes. Dengan menggunakan dataset yang sama dengan penelitian sebelumnya, yaitu dataset yang diperoleh dari situs web Kaggle, diharapkan algoritma *Random Forest* dapat menghasilkan nilai deteksi penyakit diabetes yang lebih akurat. Diharapkan juga bahwa penelitian ini akan berfungsi sebagai dasar untuk pengembangan lebih lanjut dalam ilmu data dan teknologi kesehatan.

2. METODOLOGI PENELITIAN

Pada Gambar 1 berikut menunjukkan rute penelitian: teori, hipotesis, analisis, dan rancangan.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Pertama, penelitian ini mengumpulkan data dari situs Kaggle data publik. Data pasien diabetes berjumlah 520 record dikumpulkan dari situs tersebut.

2.2 *Pre-processing*

Preprocessing data adalah proses pemilihan data sehingga dapat digunakan dengan cara yang lebih terstruktur dengan menghilangkan atribut yang tidak perlu, membuat data lebih sistematis, dan menghilangkan suara. Tahap *preprocessing*, di mana data yang hilang dan duplikat diperiksa, dikenal sebagai pembersihan data. Data yang hilang dapat diimputati dengan mean dan median atau dihapus. Data duplikat dapat dihapus[8]. Langkah pertama yaitu,

a. *Transformasi Data*

transformasi data adalah mengubah data kategori menjadi data numerik sebelum proses klasifikasi dimulai.

b. *Cleaning Data*

Cleaning Data merupakan langkah berikutnya dalam penelitian ini kumpulan data, nilainya hilang oleh penulis. Data harus dianalisis untuk menemukan informasi yang tidak relevan. Dataset memiliki nilai panjang untuk setiap variabel yang ditemukan di setiap barisnya; tidak ada nilai yang hilang.

2.3 Spliting Data

Pada tahap pembagian data, dataset yang telah diproses sebelumnya dan dibersihkan akan dibagi menjadi dua bagian. Data pelatihan akan digunakan untuk menemukan nilai kekuatan model, sedangkan data tes akan digunakan untuk mengevaluasi hasil dari nilai kekuatan model[9]. Menurut peneliti [6], Model splitting data yang digunakan untuk mempartisi dataset adalah salah satu faktor yang mempengaruhi seberapa baik model klasifikasi berfungsi pada algoritma pembelajaran mesin. Tujuan berbagi data adalah untuk memastikan bahwa model yang dibangun dapat diterapkan pada data yang baru diterima. Pada proses penelitian ini, kami memilih pembagian data terbesar, yaitu 80:20. *Random Forest* akan digunakan untuk menguji data.

2.4 Klasifikasi

Untuk membuktikan bahwa sebuah objek data termasuk dalam jenis yang telah dideskripsikan sebelumnya, klasifikasi adalah proses menemukan pola bertujuan untuk memperkirakan kelas objek yang belum diketahui[10]. Menurut peneliti [11], Salah satu metode data mining, klasifikasi menggunakan jenis analisis data yang memungkinkan prediksi berdasarkan label kelas sampel yang diklasifikasikan. Kemungkinan nilai dari atribut target menentukan kelas yang diprediksi berdasarkan nilai-nilai dari atribut prediktor; setiap instan data memiliki berbagai kemungkinan nilai[12].

2.5 Algoritma Random Forest

Random Forest (RF) adalah metode yang didasarkan pada gagasan pembelajaran kelompok, di mana beberapa pohon keputusan dibangun dan digabungkan untuk menghasilkan prediksi akhir [13]. Menurut [14], metode ini dapat meningkatkan akurasi hasil karena simpul anak dibangun secara acak untuk setiap node. 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. Nilai vektor acak yang independen dan seragam di semua pohon menentukan pembentukan algoritma ini. Saat membuat prediksi, metode classifier paling akurat digunakan. Ini memiliki kemampuan untuk menangani banyak variabel input tanpa *overfitting*, dan memiliki kemampuan untuk mengidentifikasi perbandingan antar hutan random seperti properti metode ansambel. Salah satu keuntungan menggunakan algoritma *Random Forest* sebagai metode klasifikasi adalah bahwa itu tidak menimbulkan masalah *overfitting* dalam *classifier*. *Random Forest* biasanya digunakan untuk mengidentifikasi fitur utama yang akan digunakan dari kumpulan data pelatihan dan regresi [15].

2.6 Evaluasi Kinerja Model

Evaluasi model adalah prosedur yang digunakan dalam pembelajaran mesin untuk mengevaluasi kinerja model klasifikasi. Sangat penting untuk melakukan evaluasi model untuk menilai kemampuan mereka untuk membuat klasifikasi yang akurat. Akurasi, ketepatan, recall, skor F1, dukungan, dan konsistensi matriks adalah lima komponen laporan metrix evaluasi dalam penelitian ini. Penulis dapat melihat kinerja

algoritma pembelajaran yang diawasi dengan Tabel *Confunssion Matrix*. Meskipun setiap baris memiliki contoh kelas sebenarnya, setiap kolom matiks menampilkan instance dalam kelas yang diharapkan[16].

2.6.1 Confunssion Matrix

Confunssion Matrix adalah alat yang efektif dan mudah untuk menunjukkan kinerja pengklasifikasi. Itu memiliki keuntungan karena hasilnya mudah diinterpretasikan. Kinerja model atau algoritma apa pun dapat dinilai dengan menggunakan matriks konfusi [17]. 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 1. *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 digambarkan oleh (1). Metode untuk merangkum hasil dalam matriks kebingungan yang paling umum ditunjukkan oleh (2) dan (3), masing-masing.

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

Ketika tujuannya adalah untuk mengurangi jumlah positif palsu (FP), presisi digunakan sebagai indikator kinerja. Nilai prediksi positif adalah ukuran berapa banyak dari sampel yang diprediksi positif (TP + FP) yang benar positif (TP).

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

Sensitivitas ditentukan oleh jumlah sampel positif (TP + FN) yang dianggap kelas positif (TP).

2.6.2 ROC Curve

Salah satu pendekatan yang telah dibuat untuk melakukan analisis terhadap model *classifier* adalah *kurva ROC. ROC curves* digunakan untuk menentukan parameter model yang diinginkan sesuai dengan karakteristik model *classifier*. Dengan demikian, metode klasifikasi dapat dievaluasi menggunakan standar seperti interpretabilitas, skabilitas, akurasi, kecepatan, dan kehandalan [18]. Peneliti [19], menyatakan bahwa *ROC Curve* adalah alat visual yang bermanfaat untuk membandingkan dua model klasifikasi. *Kurva Nilai ROC (Receiver Operating Characteritics)* umumnya digunakan untuk menilai hasil prediksi dalam bentuk grafik. *Kurva ROC* adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka. Nilai *AUC* dapat dibagi menjadi berbagai kelompok untuk klasifikasi data mining [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*

The Area Under Curve (AUC) dihitung untuk mengukur perbedaan performasi metode yang digunakan. *AUC* dihitung menggunakan rumus.

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

Dimana

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

Keterangan :

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

3. HASIL DAN PEMBAHASAN

Bagian ini membahas implementasi algoritma *Random Forest* dalam

mengklasifikasikan tingkat risiko penyakit diabetes. Dengan membagi data menjadi set pelatihan dan pengujian, model dibangun, diterapkan, dan diuji. Dievaluasi menggunakan *Confusin Matrix* dan *ROC AUC*.

3.1 Pengumpulan Data

Pada tahapan berikut ini dataset diperolehdarihttps://www.kaggle.com/datasets/rcratos/diabetes-risk prediction/data, dengan jumlah data 520 data yang berisikan 16 variabel dan 1 label terlihat 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* : 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.

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.

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

b. Cleaning Data

Setelah melakukan proses cleaning data pada 520 record dengan 16 atribut, kami mengkonfirmasi bahwa tidak terdapat missing value. Semua data telah diverifikasi dan siap digunakan untuk analisis selanjutnya, langkah berikutnya adalah klasifikasi.

3.3 Klasifikasi

Setelah melakukan tahapan preprocessing data dengan menggunakan *transformasi* data dan cleaning data, Selanjutnya kita melakukan tahapan klasifikasi dengan menggunakan spliting data 80:20 pada algoritma *Random Forest.*

3.3.1 Random Forest

Training model memakai fungsi *Random Forest Classifier* dari *library sklearn.ensemble. Training* data ini berguna untuk pemodelan algoritma agar 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, 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*

Pada Gambar 4, Model *Random Forest* dimulai untuk tugas klasifikasi dengan kode *rf_model = RandomForestClassifier(random_state=45)* dan mengatur biji untuk memastikan bahwa proses pengacakan dalam model dapat direplikasi. Ini adalah tahap awal dalam membangun model. Setelah itu, model akan dilatih dan dievaluasi untuk melakukan prediksi berdasarkan data yang diberikan.

Modifikasi *rf_model.Fit(X_train, y_train)* adalah bagian 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 baru di masa depan.

1.1 Evaluasi Kinerja Model

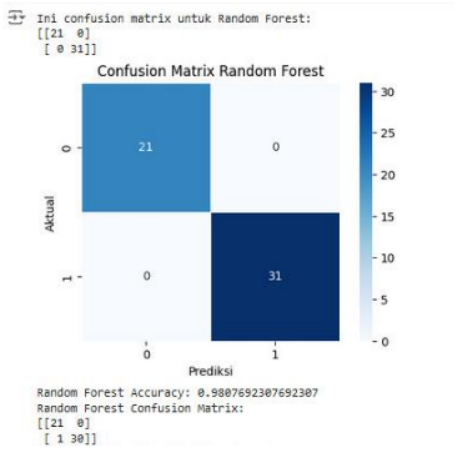
Dari hasil pengujian diatas baik evaluasi menggunakan *Confusion Matrix* maupun kurva *ROC* untuk model klasifikasi algoritma *Random Forest* sebagai berikut:

1.1.1 Confussion Matrix

Pada Tabel 2 dan Gambar 5 dibawah Merupakan hasil *Confusion Matrix* dari algoritma menggunakan *Random Forest*.

Table 2. Accurac *Random Forest*

	precision	*Recall*	*F1-score*	*support*
0	0.95	1.00	0.98	21
	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 5. *Confussion 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 :

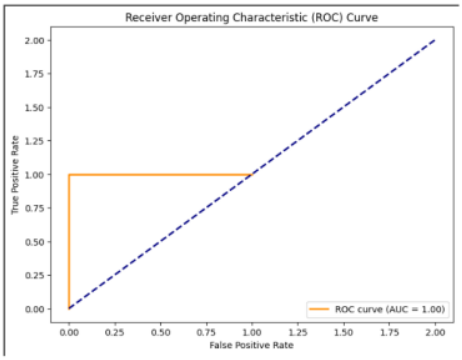
TP = 30 FP = 0
TN = 21 FN = 1

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

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

1.1.2 ROC Curve

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



Gambar 6. *ROC Random Forest*

Grafik *ROC* dengan nilai *AUC (Area Under Curve)* dengan algoritma *Random Forest* sebesar 100% dapat dilihat pada Gambar 6. Sehingga nilai *AUC* yang telah dijelaskan oleh peneliti 2 sebelumya terkait *AUC* termasuk dalam ke kategori "*Excellent Classification*" karena mempunyai nilai yang sangat baik yaitu 100%.

1.1.3 Perbandingan *Accuracy* dan *AUC*

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

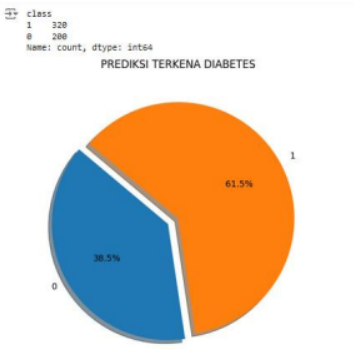
Table 3. *Hasil Accuracy dan AUC*

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

Pada Tabel 3, setelah melakukan evalusi terhadap beberapa *spliting data* penelitian ini mengambil nilai *spliting data* terbaik yaitu 80:20. hasil akurasi dan *AUC* dari metode *Random Forest* mendapatkan hasil *accuracy* 98% dan *AUC* 100%.

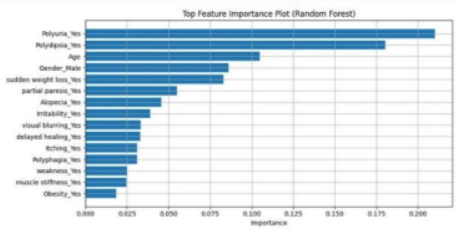
3.4 *Interprestasi* Hasil

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

Dalam penelitian ini, kami mengembangkan model klasifikasi risiko diabetes menggunakan algoritma *Random Forest*. Data yang digunakan dalam penelitian ini dibagi menjadi dua bagian dengan perbandingan 80:20 untuk pelatihan dan pengujian model. Hasil evaluasi model menunjukkan bahwa algoritma *Random Forest* dapat mengklasifikasikan risiko diabetes dengan sangat baik. Kami menemukan nilai akurasi sebesar 98%, yang menunjukkan bahwa model biasanya mampu mengklasifikasikan data dengan benar. Selain itu, nilai *Area Under Curve (AUC)* 100% menunjukkan bahwa *Area Under Curve (AUC)* di kategorikan kedalam "*Excellent Classification*" karna memiliki kemampuan yang luar biasa untuk membedakan kelas *negatif dan positif*. Temuan ini dapat membantu deteksi dan manajemen risiko diabetes secara dini, yang dapat meningkatkan kualitas hidup pasien dan mengurangi beban penyakit secara keseluruhan. Dari seleksi rangking variabel yang paling mempengaruhi pada data pasien diabetes, Hasil menunjukkan bahwa terdapat 300 pasien dengan risiko diabetes rendah dan 20 pasien dengan risiko diabetes tinggi, algoritma *Random Forest* telah ditunjukkan sebagai alat yang efektif dan dapat diandalkan untuk digunakan dalam klasifikasi risiko diabetes.

KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST

ORIGINALITY REPORT

27%
SIMILARITY INDEX

PRIMARY SOURCES

1	e-journal.stmiklombok.ac.id Internet	121 words — 4%
2	ejournal.nusamandiri.ac.id Internet	110 words — 3%
3	journal.sinov.id Internet	64 words — 2%
4	ejournal.pppmitpa.or.id Internet	62 words — 2%
5	Hendriyo Mokodompit, Nurnaningsih Nico Abdul, Elvie Fatmah Mokodongan. "PONDOK PESANTREN MODERN DARUL MADINAH WONOSARI KABUPATEN BOALEMO DENGAN PENDEKATAN ARSITEKTUR TROPIS", JAMBURA Journal of Architecture, 2024 Crossref	48 words — 1%
6	jres1.ejournal.unsri.ac.id Internet	46 words — 1%
7	docplayer.info Internet	43 words — 1%
8	ejr.stikesmuhkudus.ac.id Internet	43 words — 1%

9	digilib.unila.ac.id Internet	35 words — 1%
10	repository.ub.ac.id Internet	22 words — 1%
11	ejournal.unma.ac.id Internet	18 words — 1%
12	oaji.net Internet	18 words — 1%
13	medium.com Internet	17 words — 1%
14	www.neliti.com Internet	17 words — 1%
15	epaper.myedisi.com Internet	15 words — < 1%
16	journal.unhas.ac.id Internet	15 words — < 1%
17	www.babalawoobanifa.com Internet	14 words — < 1%
18	id.123dok.com Internet	13 words — < 1%
19	ebuah.uah.es Internet	12 words — < 1%
20	medlineplus.gov Internet	11 words — < 1%

21	123dok.com Internet	10 words — < 1%
22	jurnal.ulb.ac.id Internet	10 words — < 1%
23	www.scribd.com Internet	10 words — < 1%
24	Omar - Pahlevi, Amrin - Amrin, Yopi - Handrianto. "Implementasi Algoritma Klasifikasi Random Forest Untuk Penilaian Kelayakan Kredit", Jurnal Infortech, 2023 Crossref	9 words — < 1%
25	id.scribd.com Internet	9 words — < 1%
26	www.lokmatnews.in Internet	9 words — < 1%
27	Asrijal Bakri, Fransisco Irwandy, Elmiana Bongga Linggi. "Pengaruh Pendidikan Kesehatan Tentang Perawatan Pasien Stroke Di Rumah Terhadap Tingkat Pengetahuan Keluarga", Jurnal Ilmiah Kesehatan Sandi Husada, 2020 Crossref	8 words — < 1%
28	ejournal.bsi.ac.id Internet	8 words — < 1%
29	fk.unair.ac.id Internet	8 words — < 1%
30	hellodoktor.com Internet	8 words — < 1%
31	pesquisa.bvsalud.org	

	Internet	8 words — < 1%
32	repository.nusamandiri.ac.id Internet	8 words — < 1%
33	repository.usd.ac.id Internet	8 words — < 1%
34	siipsmarsi.wordpress.com Internet	8 words — < 1%
35	www.imic.or.jp Internet	8 words — < 1%
36	Keller. Encyclopedia of Obesity Publications	7 words — < 1%
37	ejournal.polbeng.ac.id Internet	7 words — < 1%
38	jurnal.upnyk.ac.id Internet	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF