JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 2, MEI, 2025, Pp. 223 – 234
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes

Nurussakinah [(1)*], Muhammad Faisal [(2)], Irwan Budi Santoso [(3)]
Departemen Teknik Informatika, UIN Maulana Malik Ibrahim, Malang, Indonesia
e-mail : nurussakinah2205@gmail.com, {mfaisal,irwan}@ti.uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 24 Juni 2024, direvisi 12 April 2025, diterima 15 April 2025, dan dipublikasikan 31 Mei 2025.

Abstract

Diabetes is one of the challenges in global health. Indonesia ranks 5th in the world with the highest rate of diabetes. This research uses the Random Forest algorithm for diabetes detection. The purpose of this study is to detect diabetes using the Random Forest algorithm, which provides accurate and efficient results in the early diagnosis of diabetic patients. The data used is secondary data, specifically the "Diabetes Dataset," which consists of 952 data points and has 17 features. The test scenario in this study divides the data into three parts, namely scenario 1 (90:10 ratio), scenario 2 (70:30 ratio), and scenario 3 (50:50 ratio). In each scenario, a comparison is made between using SMOTE and not using it. The best performance results are obtained in scenario 1, which uses SMOTE, producing 97% accuracy, 100% precision, 94% recall, and an F1-score of 97%.

Keywords: Detection, Diabetes, Random Forest, Synthetic Minority Oversampling Technique, Ensemble

Abstrak

Penyakit diabetes merupakan salah satu tantangan pada bidang kesehatan global. Indonesia menempati urutan ke-5 yang memiliki penyakit diabetes tertinggi dunia. Penelitian ini menggunakan algoritma Random Forest untuk deteksi diabetes. Tujuan penelitian untuk mendeteksi diabetes dengan algoritma Random Forest yang memberikan hasil akurat dan efisien dalam diagnosis dini pasien diabetes. Data yang digunakan merupakan data sekunder "*Diabetes Dataset*" yang terdiri dari 952 data dan memiliki 17 fitur. Skenario uji coba pada penelitian ini membagi data yang terdiri dari 3 bagian, yaitu skenario 1 rasio 90:10, skenario 2 rasio 70:30, skenario 3 rasio 50:50. Pada masing-masing skenario diterapkan perbandingan antara menggunakan SMOTE dan tidak. Hasil performa terbaik diperoleh pada skenario 1 yang menggunakan SMOTE yaitu, menghasilkan akurasi sebesar 97%, presisi sebesar 100%, *recall* sebesar 94%, dan yang terakhir yaitu *F1-score* yang menghasilkan 97%.

Kata Kunci: Deteksi, Diabetes, Random Forest, Synthetic Minority Oversampling Technique, *Ensemble*

1. PENDAHULUAN

Diabetes adalah penyakit kronis yang berdampak signifikan pada kesehatan global (Sun et al., 2023). Berdasarkan data International Diabetes Federation (IDF) tahun 2021, sekitar 537 juta orang berusia 20-79 tahun menderita diabetes (Magliano & Boyko, 2013). Indonesia menempati peringkat kelima di dunia dalam jumlah kasus tertinggi yang mencapai sekitar 19,5 juta jiwa di tahun 2021 dan hal ini akan terus meningkat setipa tahunnya yang diperkirakan mencapai angka 28,6 juta di tahun 2045. Kementerian Kesehatan Republik Indonesia (2023) menyatakan bahwa diabetes menjadi angka kematian tertinggi ketiga setelah penyakit strok dan jantung.

Diabetes ditandai dengan hiperglikemia, yaitu kondisi kadar gula darah yang tinggi akibat ketidakmampuan tubuh untuk memproduksi atau menggunakan insulin dengan efektif (Hana, 2020). Insulin sebagai hormon yang penting dalam tubuh untuk mengatur metabolisme gula darah (Rahman et al., 2021). Berbagai faktor lain seperti gaya hidup dan faktor genetik juga berperan dalam risiko mengalami diabetes (Faida & Santik, 2020). Pentingnya diagnosis dini

sangat ditekankan untuk mencegah komplikasi serius dan meningkatkan kualitas hidup penderita (Karyadiputra & Setiawan, 2022).

Kemajuan teknologi khususnya di bidang *data mining* memiliki potensi yang sangat besar untuk mengidentifikasi pola data yang penting dalam diagnosis dini diabetes (Elfaladonna & Rahmadani, 2019). Salah satu metode *data mining* yang efektif adalah deteksi. Deteksi memungkinkan pengelompokan data pasien berdasarkan fitur-fitur tertentu (Aris & Benyamin, 2019). Algoritma Random Forest merupakan bagian dari metode *decision tree*, yang dikenal memiliki akurasi tinggi dan kemampuannya dalam menangani data dalam jumlah besar (Li & Mu, 2024).

Penelitian terkait tentang klasifikasi dan deteksi diabetes telah berkembang dengan berbagai pendekatan dan algoritma. Witjaksana et al. (2021) membandingkan akurasi klasifikasi diabetes menggunakan algoritma Random Forest dan Artificial Neural Network (ANN). Hasil menunjukkan bahwa Random Forest memiliki akurasi sebesar 90,62%, sedangkan ANN memiliki akurasi sebesar 82,29%. Hal ini menegaskan bahwa Random Forest adalah algoritma yang lebih efektif dalam klasifikasi diabetes dibandingkan ANN.

Penelitian oleh Daghistani & Alshammari (2020) membandingkan prediksi diabetes menggunakan Logistic Regression dan Random Forest. Hasil penelitian menunjukkan bahwa model Random Forest memiliki presisi 0,883, *recall* 0,88, dan *F-Measure* 0,876, sedangkan Logistic Regression memiliki presisi 0,692, *recall* 0,703, dan *F-Measure* 0,675. AUC untuk Random Forest adalah 0,944, dibandingkan Logistic Regression yang memiliki AUC sebesar 0,708. Penelitian ini menegaskan bahwa Random Forest memiliki kinerja prediksi yang lebih unggul dibandingkan Logistic Regression dalam mendeteksi diabetes.

Sistematis *review* yang dilakukan oleh Tulu et al. (2023) lebih lanjut meneliti keunggulan SMOTE dalam berbagai aplikasi pembelajaran mesin. Seperti halnya klasifikasi dan prediksi kesehatan, dengan menekankan bahwa teknik ini secara konsisten meningkatkan kinerja model dalam menghadapi *dataset* yang tidak seimbang.

Penelitian ini mengimplementasikan algoritma Random Forest dengan Synthetic Minority Oversampling Technique (SMOTE) untuk mendeteksi diabetes. SMOTE digunakan untuk menyeimbangkan data dengan cara *resampling*, sehingga model dapat mengidentifikasi pasien yang berisiko tinggi secara lebih akurat. Penelitian ini berupaya untuk menyatukan kedua aspek tersebut, memberikan pendekatan holistik yang dapat diterapkan dalam praktek medis untuk meningkatkan deteksi dini dan penanganan diabetes. Dengan diagnosis dini yang lebih tepat, intervensi awal seperti pemantauan khusus dan penyesuaian pola hidup dapat dilakukan untuk mencegah perkembangan diabetes dan komplikasi yang lebih serius.

2. METODE PENELITIAN

2.1 *Dataset* Penelitian

Data pada penelitian ini merupakan Diabetes Dataset yang berasal *Birla Institute of Technology* (Tigga & Garg, 2020). *Dataset* ini kumpulan data yang dikumpulkan oleh Tigga & Garg (2020) dari Jurusan Ilmu dan Teknik Komputer. *Dataset* pada Tabel 1 berjumlah 952 data yang terdiri dari 17 fitur dan 1 kelas target. Fitur '*Diabetic*' merupakan kelas target yang menunjukkan bahwa itu termasuk diabetes atau bukan diabetes.

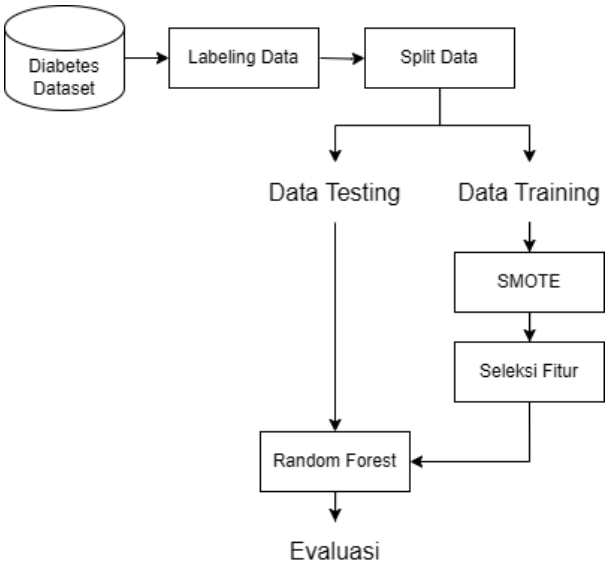
Tabel 1 Fitur Diabetes *Dataset*

No.	Fitur	Deskripsi
1	*Age*	Usia pasien
2	*Gender*	Jenis kelamin pasien
3	*Family Diabetes*	Riwayat keluarga dengan penyakit diabetes
4	*HighBP*	Tekanan darah tinggi pasien
5	*PhysicallyActive*	Aktivitas sehari-hari / aktivitas bergerak
6	BMI	Indeks Massa Tubuh
7	*Smoking*	Pasien termasuk dari perokok atau tidak
8	*Alcohol*	Mengkonsumsi alkohol atau tidak
9	*Sleep*	Jam tidur pasien
10	*SoundSleep*	Jam tidur nyenyak
11	*RegularMedicine*	Asupan obat secara teratur
12	*JunkFood*	Mengkonsumsi junk food atau tidak
13	*Strees*	Berapa tingkat mengalami stres
14	*BPLevel*	Tingkat tekanan darah tinggi
15	*Pregnancies*	Kehamilan
16	*Pdiabetes*	Kadar gula darah pada wanita hamil
17	*UrinationFreq*	Frekuensi Urine
18	*Diabetic*	Merupakan diabetes atau tidak (kelas target)

2.2 Desain Sistem

Pada penelitian ini terdapat desain sistem yang merupakan tahapan-tahapan yang dilakukan mulai dari *input* hingga menghasilkan hasil evaluasi berupa akurasi, *recall,* presisi, dan *F1-score*. Desain sistem pada Gambar 1 menunjukkan tahap-tahap pembangunan model yang dijabarkan sebagai berikut:

a) *Input* diabetes *dataset*.
b) Pengubahan data dengan cara *ordinal encoding* untuk merubah fitur kategorik menjadi numerik guna untuk mempermudah tahap pemodelan. Tahap ini dilakukan proses pengubahan data untuk memastikan *dataset* siap untuk digunakan dalam evaluasi model. Dataset yang memiliki fitur kategorik perlu dikonversikan menjadi numerik supaya dapat diolah dengan mudah oleh model.
c) Pembagian data menjadi tiga pengujian skenario. Skenario 1 membagi data menjadi 90:10, skenario 2 data dibagi menjadi 70:30 dan skenario 3 adalah 50:50. Pembagian 3 skenario digunakan untuk mengevaluasi model pada kondisi yang berbeda-beda. Skenario 1 melatih model dengan data latih yang jauh lebih besar daripada data uji untuk melihat kemampuan generalisasi meskipun data uji yang terbatas. Skenario 2 yang umum digunakan pada pembagian data latih dan data uji sehingga data ini dipakai untuk perbandingan dengan skenario yang lainnya. Pada skenario 3, mengevaluasi model dengan data latih dan data uji yang seimbang, apakah hal ini mempengaruhi performa model atau tidak. Sehingga dengan ketiga skenario tersebut perlu untuk dilakukan supaya dapat mengetahui mana skenario yang menghasilkan evaluasi model terbaik.
d) SMOTE untuk menyeimbangkan *dataset* yang telah melalui tahap split data dengan menggunakan data train saja.
e) Seleksi fitur dilakukan untuk memilih fitur yang paling relevan dalam pemodelan untuk meningkatkan performa model (Mostafa et al., 2024). Pada penelitian ini teknik seleksi fitur yang digunakan adalah *Feature Importance* dari algoritma Random Forest *classifier*. Teknik ini dapat mengevaluasi masing-masing fitur terhadap hasil prediksi.
f) Kemudian implementasi algoritma Random Forest pada penelitian ini untuk membangun model yang menghasilkan nilai performa model dari *confusion matrix.* Evaluasi model yang paling baik dilihat pada hasil akurasi, presisi, *recall* dan *F1-score* dari masing-masing skenario dan pengujian dengan menggunakan SMOTE dan tanpa SMOTE.



Gambar 1 Desain Penelitian

2.3 Random Forest

Random Forest merupakan algoritma pembelajaran mesin berbasis *ensemble* yang terdiri dari sejumlah pohon keputusan atau yang biasa disebut dengan *decision trees* yang bekerja bersama untuk meningkatkan performa model (Zailani & Hanun, 2020). Algoritma ini terkenal karena kemampuannya dalam menangani *dataset* besar dan kompleks, serta mengurangi risiko *overfitting* yang umum pada pohon keputusan tunggal (Junus et al., 2023). Beberapa kelebihan dari Random Forest selain mampu mengatasi *overfitting* juga toleransi pada data tidak eimbang dan mampu menangani data yang *missing* (Rajaraman & Ullman, 2011). Langkah-langkah pembuatan Random Forest sebagai berikut:

1) Penenutan jumlah pohon yang akan dibuat.
2) Membangun *tree* dari setiap sampel data dengan banyaknya sesuai dengan yang ditentukan.
3) Membuat *tree* dari kumpulan pohon-pohon. Kemudian mengulangi langkah 1 dan 2 hingga mencapai jumlah jumlah pohon yang telah ditentukan sebelumnya.
4) Hasil akhir yaitu dengan *majority vote* dari hasil semua pohon keputusan.

2.4 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE adalah salah satu teknik *oversampling* yang digunakan untuk menambah data minoritas pada dataset yang tidak seimbang (Yang et al., 2024). SMOTE bekerja untuk menghasilkan sampel sintesis baru pada data minoritas (Mulia & Kurniasih, 2023). Data sintesis dibuat berdasarkan prinsip kerja K-Nearest Neighbor (tetangga terdekat). Langkah SMOTE dimulai dengan identifikasi data minoritas dan menentukan k tetangga terdekat dengan persamaan jarak Euclidean pada Pers. (1). Di mana pada persamaan tersebut $d(x, y)$ merupakan jarak Euclidean, xi dan yi adalah koordinat x dan y pada dimensi ke-i, dan n menunjukkan jumlah dimensi Euclidean.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad (1)$$

Selanjutnya, SMOTE melakukan pemilihan acak dari k tetangga dan menghitung perbedaan antara fitur minoritas asli dan fitur tetangga yang dipilih. mengalikan perbedaan tersebut dengan bilangan acak, dan menambahkannya ke sampel asli untuk menghasilkan sampel sintesis.

Sampel sintesis ini kemudian ditambahkan ke *dataset*, meningkatkan jumlah sampel dalam kelas minoritas dan membantu model pembelajaran mesin belajar dari data yang lebih seimbang.

3. HASIL DAN PEMBAHASAN

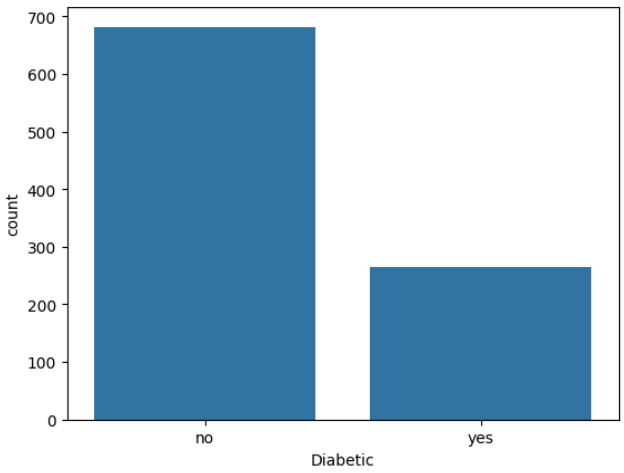
Penelitian ini memiliki beberapa tahap yang akan dibahas antara lain, pengubahan data, SMOTE, seleksi fitur, pemodelan serta evaluasi model. Setiap tahapan memiliki peran penting dalam membangun model klasifikasi yang optimal. Sebelum ke tahapan tersebut, berikut pada Tabel 9 di Lampiran A merupakan lima contoh dari diabetes *dataset*.

3.1 Pengubahan Data

Dataset pada Tabel 10 di Lampiran A selanjutnya dilakukan tahap pengubahan data. Teknik yang dilakukan pada tahap ini adalah *ordinal encoder* di mana merubah ketegorik menjadi numerik yang memiliki tingkatan seperti pada fitur '*Age*'. Dapat dilihat hasil setelah pengubahan dari *dataset*, data yang memiliki tingkatan akan diurutkan mulai dari 0. '*Age'* memiliki 4 tingkatan yaitu usia 40-49, 50-59, usia lebih dari 60 dan kurang dari 40. Dengan teknik *ordinal*, maka tingkatan usia tersebut dirubah menjadi numerik yaitu, mulai dari 0, 1, 2 dan 3.

3.2 SMOTE

Diabetes *dataset* memiliki data yang tidak seimbang pada kelas target yang merupakan diabetes dan bukan diabetes. Sehingga dilakukan teknik *resampling* yaitu SMOTE untuk menyeimbangkan *dataset* tersebut. Gambar 2 jika dilihat antara kelas bukan diabetes dan diabetes memiliki jumlah yang sangat berbeda. Kelas target bukan diabetes ada sekitar 700 data dan kelas target diabetes sekitar 250 data. Pada tahap ini dilakukan setelah split data, dan data yang digunakan pada tahap SMOTE hanya pada data *training* saja supaya tidak terjadi kebocoran data yang mempengaruhi hasil pada performa model.



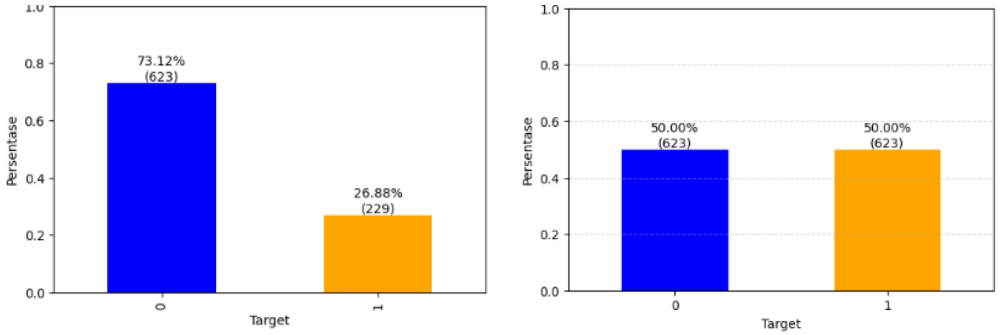
Gambar 2 Perseberan Data pada Kelas Target

Pada skenario 1 data dibagi menjadi 90% data *train* dan 10% data *test*. Data *train* yang dihasilkan pada kelas bukan diabetes adalah 623 data sedangkan data diabetes sebanyak 229 data pada Gambar 3(a). Maka, penyeimbangan data dilakukan yang kemudian menghasilkan jumlah yang sama antara bukan diabetes dan diabetes. Setelah dilakukan SMOTE, data pada kelas diabetes yang awalnya jauh berbeda dengan data bukan diabetes menjadi sama yaitu dengan jumlah 623 pada data diabetes yang bisa dilihat di Gambar 3(b).

Pada skenario 2 data dibagi menjadi 70% data *train* dan 30% data *test*. Data *train* yang dihasilkan pada kelas bukan diabetes yaitu 483 data dan kelas diabetes sebanyak 179 seperti pada Gambar 4(a). Kemudian menghasilkan jumlah yang sama antara bukan diabetes dan diabetes setelah

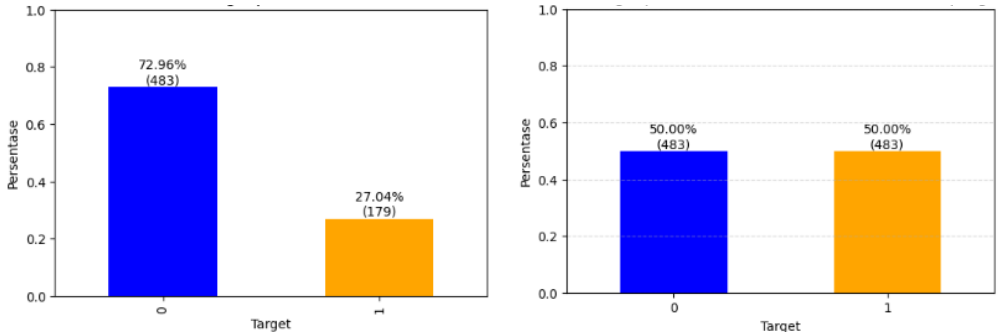
dilakukan SMOTE. Data pada kelas diabetes menjadi sama dengan data bukan diabetes yaitu menjadi berjumlah 483 pada data diabetes yang bisa dilihat di Gambar 4(b).

Pada skenario 3 data dibagi menjadi 50% data *train* dan 50% data *test*. Data *train* yang dihasilkan pada kelas bukan diabetes sebanyak 343 data dan pada kelas diabetes sebanyak 130 data seperti pada Gambar 5(a). Setelah itu, melakukan penyeimbangan data yang menghasilkan jumlah yang sama antara kelas bukan diabetes dan kelas diabetes. Setelah SMOTE, kelas diabetes menjadi 343 yang bisa dilihat di Gambar 5(b).



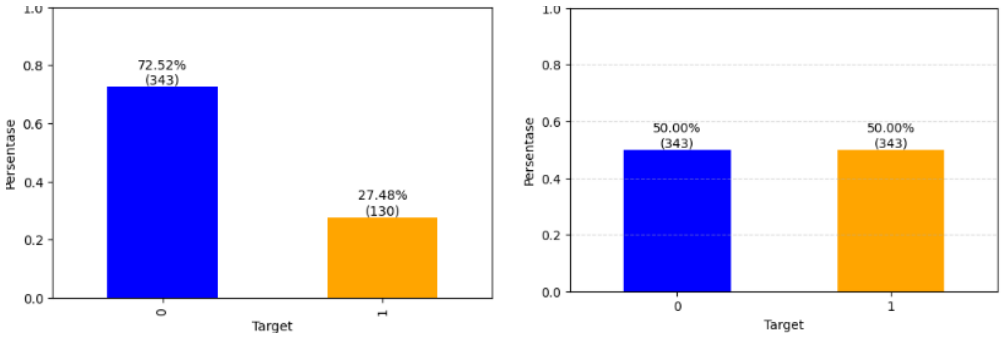
(a) (b)

Gambar 3 Sebelum (a) dan Setelah (b) SMOTE 90:10



(a) (b)

Gambar 4 Sebelum (a) dan Setelah (b) SMOTE 70:30



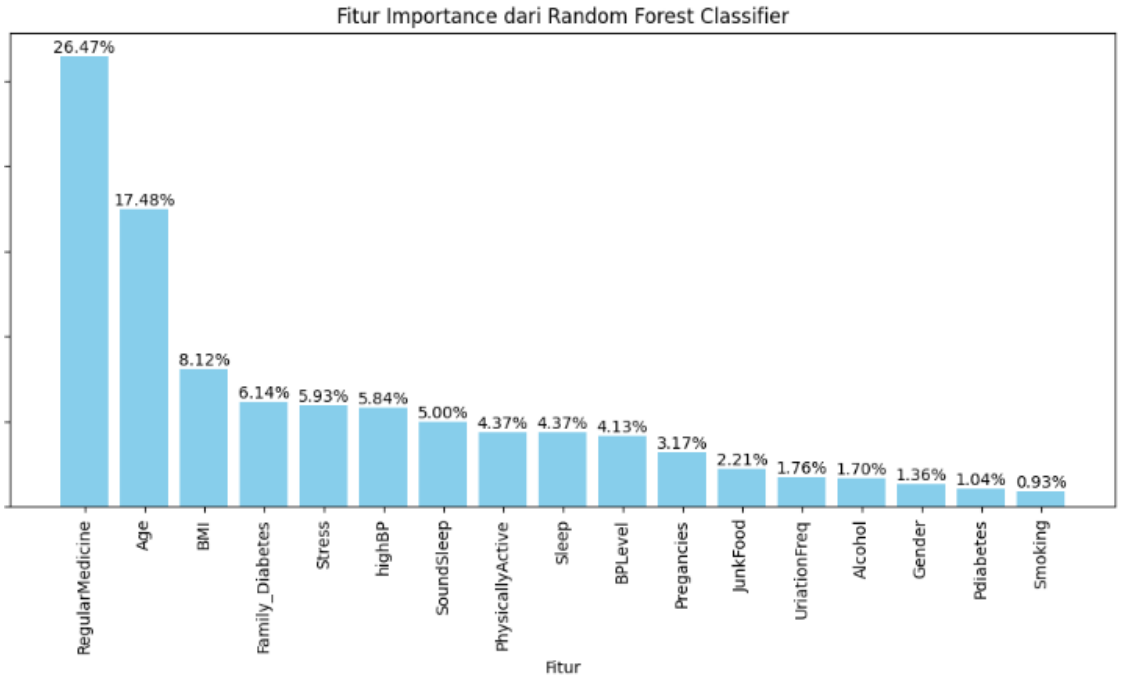
(a) (b)

Gambar 5 Sebelum (a) dan Setelah (b) SMOTE 50:50

3.3 Seleksi Fitur

Seleksi fitur dilakukan untuk memilih fitur yang relevan pada *dataset* untuk dilakukan pemodelan. *Dataset* yang digunakan oleh peneliti memiliki fitur yang berjumlah 17. Dari seluruh fitur tersebut dilakukan seleksi untuk mengambil fitur yang paling relevan yaitu 10 dari 17 fitur yang telah

diseleksi. Hasil dari 17 fitur pada *dataset*, dilakukan seleksi dengan teknik *Feature Importance* dari Random Forest. Hasil dari seleksi menunjukkan ada 10 fitur yang paling relevan antara lain, *regular medicine, age,* BMI*, family diabetes, stress, highbp, soundsleep, physically active, sleep*, dan *bplevel* seperti yang ditunjukkan pada Gambar 6. Dari Tabel 2, fitur *regular medicine, age,* dan BMI memiliki pengaruh terbesar terhadap hasil model, sedangkan fitur lain tetap relevan dengan bobot lebih kecil. Fitur-fitur yang memiliki hasil *importance score* kurang dari 4,00% diabaikan untuk meningkatkan efisiensi model.



Gambar 6 Grafik Hasil Seleksi Fitur

Tabel 2 Hasil Seleksi Fitur

No.	Fitur	*Importance Score*
1	Regular Medicine	26,47%
2	Age	17,48%
3	BMI	8,12%
4	Family_Diabetes	6,14%
5	Stress	5,93%
6	HighBP	5,84%
7	SoundSleep	5,00%
8	Physically Active	4,37%
9	Sleep	4,37%
10	BPLevel	4,13%

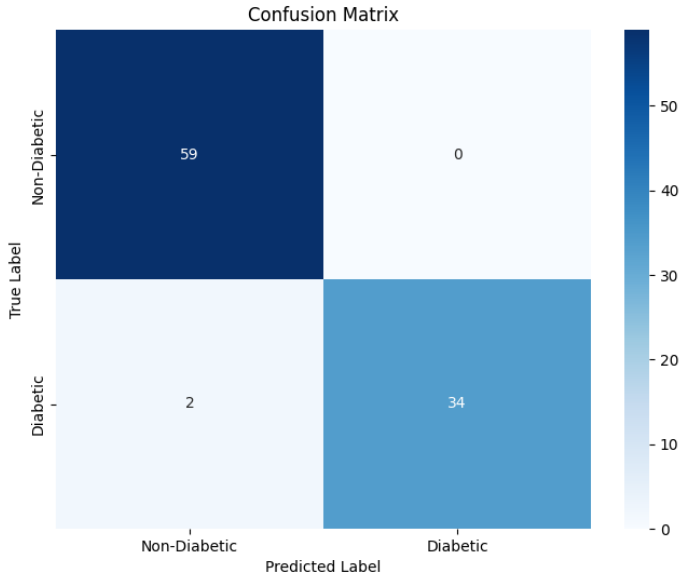
3.4 Evaluasi Random Forest dengan SMOTE

Terdapat tiga skenario yang dilakukan pada penelitian ini, yaitu membagi data latih dan data uji menjadi rasio antara lain 90:10, 70:30, 50:50. Setiap skenario digunakan untuk menguji kestabilan dan performa model dalam melakukan klasifikasi. Hasil evaluasi model dengan algoritma Random Forest sebagai berikut. Setelah dilakukan pembagian data pada skenario 1 dengan rasio 90:10, selanjutnya melakukan penyeimbangan data pada data *training*. Kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik. Setelah itu dilakukan *tuning* parameter untuk mendapatkan *hyperparameter* terbaik, yaitu *max_depth* sebesar 50 dan *n_estimators* sebanyak 10. Skenario 1 dengan tahap SMOTE ini, menghasilkan *confusion matrix*

pada Tabel 3 dengan visualisasi grafik pada Gambar 7. Pada tabel dan gambar tersebut memperlihatkan bahwa ada 34 yang benar diprediksi diabetes dan 59 yang diprediksi dengan benar bukan diabetes. Ada 2 yang merupakan diabetes tetapi diprediksi bukan diabetes. Dari *confusion* matrix tersebut menghasilkan evaluasi model berupa akurasi sebesar 97%, presisi 100%, *recall* 94%, dan *F1-score* sebesar 97%.

Tabel 3 *Confusion Matrix* Skenario 1 dengan SMOTE

Prediksi	Aktual	
	Diabetes	Bukan Diabetes
Diabetes	34	0
Bukan Diabetes	2	59



Gambar 7 Grafik *Confusion Matrix* Skenario 1 dengan SMOTE

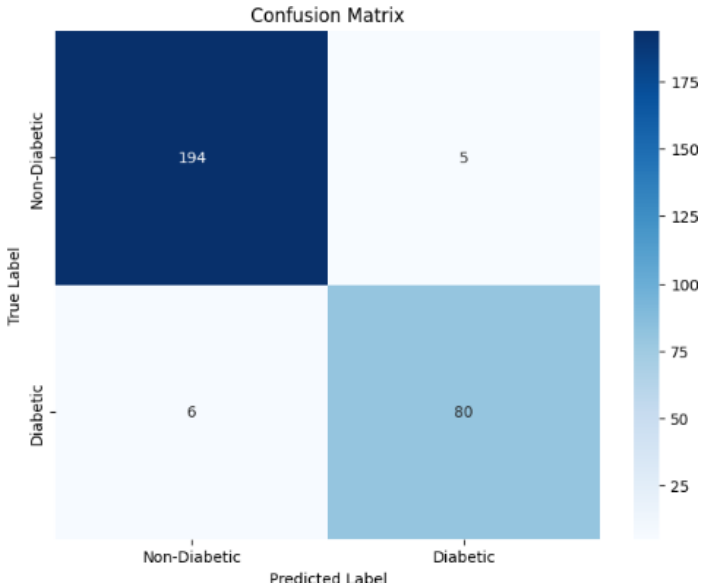
Setelah melakukan pembagian data (*split*) dan penyeimbangan data pada data *training* menggunakan skenario 2 dengan rasio 70:30, kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik. Selanjutnya dilakukan *tuning* parameter untuk mendapatkan *hyperparameter* terbaik, yaitu *max_depth* sebesar 100 dan *n_estimators* sebanyak 20. Skenario 2 dengan tahap SMOTE menghasilkan *confusion matrix* pada Tabel 4 dengan visualisasi *confusion matrix* pada Gambar 8. Hasil tersebut memperlihatkan bahwa ada 80 yang diprediksi secara benar diabetes dan 194 bukan diabetes. Ada 6 yang diabetes tetapi diprediksi bukan diabetes dan 5 bukan diabetes diprediksi sebagai diabetes. Dari hasil *confusion matrix*, evaluasi model menghasilkan akurasi sebesar 96%, presisi 94%, *recall* 93%, dan *F1-score* sebesar 94%.

Setelah dilakukan pembagian data (*split*) pada skenario 3 dengan rasio 50:50, selanjutnya menerapkan SMOTE pada data *training*. Kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik untuk pemodelan. Selanjutnya melakukan tuning parameter untuk mendapatkan *hyperparameter* terbaik, yaitu *max_depth* sebesar 50 dan *n_estimators* sebanyak 15. Skenario 3 yang menggunakan SMOTE, hasil evaluasi ditampilkan dalam *confusion matrix* pada Tabel 5.

Gambar 9 merupakan hassil *confusion matrix* skenario 3 dengan SMOTE yang menunjukkan ada 121 data yang diprediksi secara benar diabetes dan 327 bukan diabetes. Ada 14 data diabetes tetapi diprediksi bukan diabetes dan 12 data bukan diabetes diprediksi diabetes. Dari data tersebut, evaluasi model menghasilkan akurasi sebesar 95%, presisi 91%, *recall* 90%, dan *F1-score* sebesar 90%.

Tabel 4 *Confusion Matrix* Skenario 2 dengan SMOTE

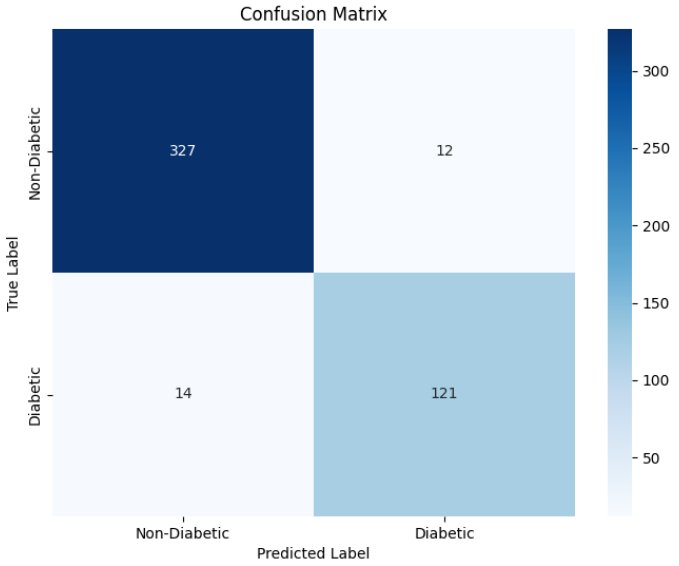
Prediksi	Aktual	
	Diabetes	Bukan Diabetes
Diabetes	80	5
Bukan Diabetes	6	194



Gambar 8 Grafik *Confusion Matrix* Skenario 2 dengan SMOTE

Tabel 5 *Confusion Matrix* Skenario 3 dengan SMOTE

Prediksi	Aktual	
	Diabetes	Bukan Diabetes
Diabetes	121	12
Bukan Diabetes	14	327



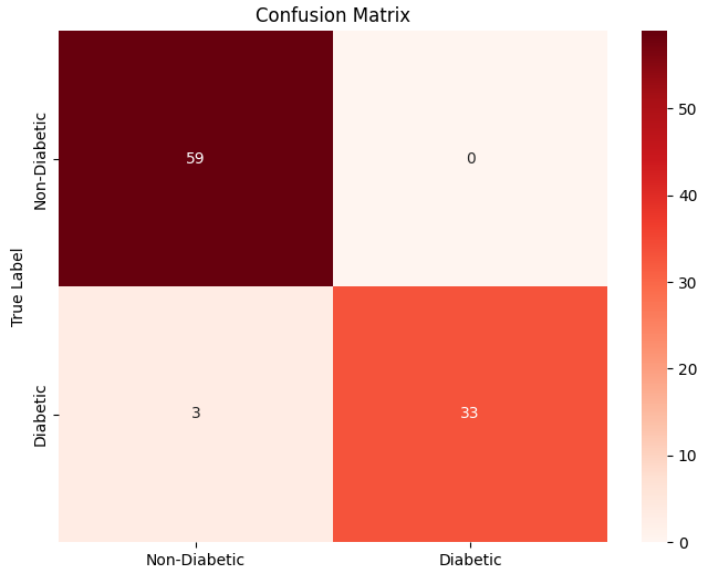
Gambar 9 Grafik *Confusion Matrix* Skenario 3 dengan SMOTE

3.5 Evaluasi Random Forest Tanpa SMOTE

Setelah dilakukan pembagian data pada skenario 1 dengan rasio 90:10, selanjutnya melakukan penyeimbangan data pada data training. Kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik untuk pemodelan. Selanjutnya dilakukan *tuning* parameter untuk menemukan *hyperparameter* terbaik, yaitu *max_depth* sebesar 50 dan *n_estimators* sebanyak 10. Skenario 1 tanpa SMOTE ini menghasilkan *confusion matrix* pada Tabel 6 dengan visualisasi pada Gambar 10. Hasil *confusion matrix* tersebut menunjukkan bahwa ada 33 yang benar diprediksi diabetes dan 59 yang diprediksi dengan benar bukan diabetes. Ada 3 yang merupakan diabetes tetapi diprediksi bukan diabetes. Maka skenario ini menghasilkan evaluasi model berupa akurasi sebesar 96%, presisi 100%, *recall* 92%, dan *F1-score* sebesar 96%.

Tabel 6 *Confusion Matrix* Skenario 1 Tanpa SMOTE

Prediksi	Aktual	
	Diabetes	Bukan Diabetes
Diabetes	33	0
Bukan Diabetes	3	59

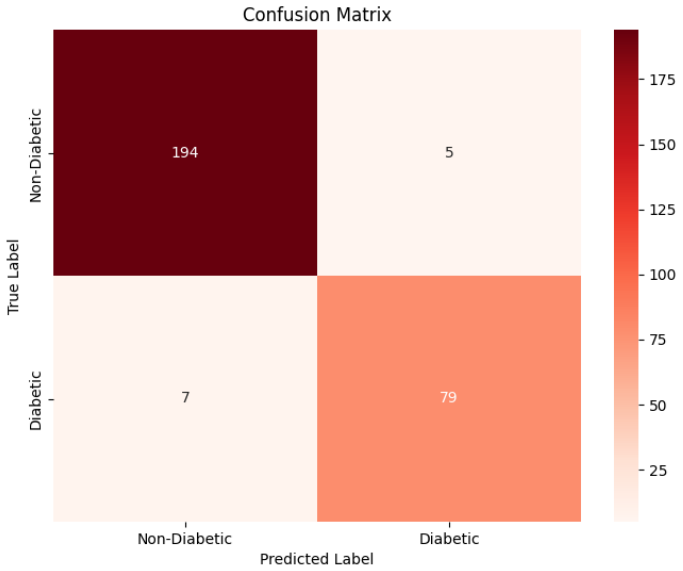


Gambar 10 Grafik *Confusion Matrix* Skenario 1 Tanpa SMOTE

Setelah dilakukan pembagian data pada skenario 2 dengan rasio 70:30, selanjutnya melakukan penyeimbangan data pada data *training*. Kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik untuk pemodelan. Selanjutnya dilakukan *tuning* parameter untuk mendapatkan *hyperparameter* terbaik, yaitu *max_depth* sebesar 50 dan *n_estimators* sebanyak 10. Skenario 2 tanpa tahap SMOTE menghasilkan *confusion matrix* pada Tabel 7 dengan grafik pada Gambar 11. Hasil *confusion matrix* tersebut memperlihatkan bahwa ada 79 yang diprediksi secara benar diabetes dan 194 bukan diabetes. Ada 7 yang diabetes tetapi diprediksi bukan diabetes dan 5 bukan diabetes diprediksi sebagai diabetes. Dari hasil tersebut maka evaluasi model menghasilkan akurasi sebesar 96%, presisi 94%, *recall* 92%, dan *F1-score* sebesar 93%.

Tabel 7 *Confusion Matrix* Skenario 2 Tanpa SMOTE

Prediksi	Aktual	
	Diabetes	Bukan Diabetes
Diabetes	79	5
Bukan Diabetes	7	194

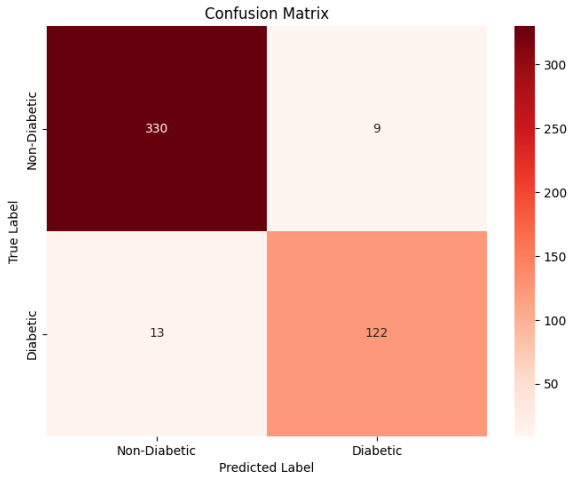


Gambar 11 Grafik *Confusion Matrix* Skenario 2 Tanpa SMOTE

Setelah dilakukan pembagian data pada skenario 1 dengan rasio 50:50, lalu dilakukan penerapan SMOTE pada data *training*. Kemudian dilakukan seleksi fitur dengan menggunakan 10 fitur terbaik. Selanjutnya melakukan tuning parameter untuk mendapatkan *hyperparameter* terbaik, yaitu *max_depth* sebesar 50 dan *n_estimators* sebanyak 10. Skenario 3 tanpa SMOTE menghasilkan *confusion matrix* pada Tabel 8 dengan visualisasi pada Gambar 12.

Tabel 8 *Confusion Matrix* Skenario 3 Tanpa SMOTE

Prediksi	Aktual	
	Diabetes	**Bukan Diabetes**
Diabetes	122	9
Bukan Diabetes	13	330



Gambar 12 Grafik *Confusion Matrix* Skenario 3 Tanpa SMOTE

Hasil *confusion matrix* skenario 3 tanpa SMOTE yang menunjukkan ada 122 data yang diprediksi secara benar diabetes dan 330 bukan diabetes. Ada 13 data diabetes tetapi diprediksi bukan diabetes dan 9 data bukan diabetes diprediksi diabetes. Dari data tersebut, evaluasi model menghasilkan akurasi sebesar 95%, presisi 93%, *recall* 90%, dan *F1-score* sebesar 92%.

Dari hasil evaluasi model dengan Random Forest yang telah terbagi dalam tiga skenario, hasil paling baik yaitu pada skenaro 1 yang merupakan membagi rasio data 90% data latih dan 10% data uji. Skenario 1 memiliki data latih yang lebih besar daripada skenario 2 dan 3. Maka pembagian rasio data berpengaruh pada bagaimana model dilatih. Pembagian data latih yang lebih banyak dari data uji cenderung menghasilkan performa model yang lebih baik. Data latih digunakan untuk melatih model. Model yang memiliki informasi pada data latih yang lebih banyak menyebabkan model mampu mempelajari pola data dengan lebih kompleks (Barus, 2022). Dengan demikian, model dapat generalisasi pada data baru dan menghasilkan model yang lebih akurat.

Evaluasi dengan SMOTE juga menghasilkan hasil yang lebih baik daripada tanpa SMOTE. Penggunaan SMOTE dapat meningkatkan performa model pada data yang tidak seimbang dengan menyeimbangkan jumlah sampel antara kelas mayoritas dan minoritas. Tanpa SMOTE, model cenderung hanya memperhatikan pada kelas mayoritas saja dan mengabaikan kelas minoritas (Sutoyo et al., 2020). Dengan SMOTE, lebih banyak contoh dari kelas minoritas yang dihasilkan secara sintetis, serta membantu model untuk lebih baik dalam mengenali dan memprediksi kedua kelas target sehingga dapat meningkatkan performa model.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan metode deteksi dini diabetes menggunakan algoritma Random Forest yang dikombinasikan dengan teknik SMOTE untuk menyeimbangkan data. Hasil menunjukkan bahwa kualitas model prediksi sangat dipengaruhi oleh proporsi data latih dan keseimbangan distribusi kelas target. Skenario terbaik adalah skenario 1 dengan pembagian data 90% untuk latih dan 10% untuk uji, yang menghasilkan kinerja tertinggi dengan akurasi 97%, presisi 100%, *recall* 94%, dan F1-*score* 97%. Ini membuktikan bahwa data latih yang lebih banyak dan seimbang memungkinkan model belajar lebih efektif dalam mengenali tanda-tanda awal diabetes.

Penelitian ini juga terdapat pemilihan fitur menggunakan *Feature Importance* dari *Random Forest* dan penerapan SMOTE dalam berbagai skenario pembagian data. Hal ini memberikan dasar penting bagi pengembangan sistem deteksi dini penyakit diabetes, khususnya untuk data penyakit dengan ketidakseimbangan data yang tinggi seperti diabetes. Penelitian selanjutnya bisa menggunakan model lain, seperti *cross-validation* dan menambahkan data klinis untuk meningkatkan kemampuan model pada kasus nyata.

DAFTAR PUSTAKA

Aris, F., & Benyamin, B. (2019). Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi. *Router Research*, *1*(1), 1–6. https://doi.org/10.29239/j.router.2019.313

Daghistani, T., & Alshammari, R. (2020). Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes. *Journal of Advances in Information Technology*, *11*(2), 78–83. https://doi.org/10.12720/jait.11.2.78-83

Elfaladonna, F., & Rahmadani, A. (2019). Analisa Metode Classification-Decision Tree dan Algoritma C.45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner. *SINTECH (Science and Information Technology) Journal*, *2*(1), 10–17. https://doi.org/10.31598/sintechjournal.v2i1.293

Faida, A. N., & Santik, Y. D. P. (2020). Kejadian Diabetes Melitus Tipe I pada Usia 10-30 Tahun. *Higeia Journal of Public Health Research and Development*, *4*(1), 33–42. https://doi.org/10.15294/higeia/v4i1/31763

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, *4*(1), 32–39. https://doi.org/10.47970/siskom-kb.v4i1.173

Junus, C. Z. V., Tarno, T., & Kartikasari, P. (2023). Klasifikasi Menggunakan Metode Support Vector Machine dan Random Forest untuk Deteksi Awal Risiko Diabetes Melitus. *Jurnal Gaussian*, *11*(3), 386–396. https://doi.org/10.14710/j.gauss.11.3.386-396

Karyadiputra, E., & Setiawan, A. (2022). Penerapan Data Mining untuk Prediksi Awal Kemungkinan Terindikasi Diabetes. *Teknosains: Media Informasi Sains dan Teknologi*, *16*(2), 221–232. https://doi.org/10.24252/teknosains.v16i2.28257

Kementerian Kesehatan Republik Indonesia. (2023). *Rencana Aksi Kerja Kegiatan Direktorat P2PTM 2021-2024* (1st ed.). Kementerian Kesehatan Republik Indonesia. https://www.scribd.com/document/757455987/RAK-Dit-P2PTM-1-465827-02-4tahunan-070

Li, Y., & Mu, Y. (2024). Research and Performance Analysis of Random Forest-Based Feature Selection Algorithm in Sports Effectiveness Evaluation. *Scientific Reports*, *14*(1), Article ID: 26275. https://doi.org/10.1038/s41598-024-76706-1

Magliano, D., & Boyko, E. J. (2013). Five Questions on the IDF Diabetes Atlas. *Diabetes Research and Clinical Practice*, *102*(2), 147–148. https://doi.org/10.1016/j.diabres.2013.10.013

Mostafa, G., Mahmoud, H., Abd El-Hafeez, T., & ElAraby, M. E. (2024). The Power of Deep Learning in Simplifying Feature Selection for Hepatocellular Carcinoma: A Review. *BMC Medical Informatics and Decision Making*, *24*(1), Article ID: 287. https://doi.org/10.1186/s12911-024-02682-1

Mulia, C., & Kurniasih, A. (2023). Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Bank Customer Churn Menggunakan Algoritma Naïve bayes dan Logistic Regression. *Prosiding Seminar Ilmiah Nasional Online Mahasiswa Ilmu Komputer dan Aplikasinya*, *4*(2), 552–559. https://conference.upnvj.ac.id/index.php/senamika/article/view/2590

Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, Md. A., Hannan, Md. A., Uddin, M. J., & Pang, M.-G. (2021). Role of Insulin in Health and Disease: An Update. *International Journal of Molecular Sciences*, *22*(12), Article ID: 6403. https://doi.org/10.3390/ijms22126403

Rajaraman, A., & Ullman, J. D. (2011). Data Mining. In *Mining of Massive Datasets* (Vol. 2, Issue January 2013, pp. 1–17). Cambridge University Press. https://doi.org/10.1017/CBO9781139058452.002

Sun, J., Hu, W., Ye, S., Deng, D., & Chen, M. (2023). The Description and Prediction of Incidence, Prevalence, Mortality, Disability-Adjusted Life Years Cases, and Corresponding Age-Standardized Rates for Global Diabetes. *Journal of Epidemiology and Global Health*, *13*(3), 566–576. https://doi.org/10.1007/s44197-023-00138-9

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes Using Machine Learning Classification Methods. *Procedia Computer Science*, *167*, 706–716. https://doi.org/10.1016/j.procs.2020.03.336

Tulu, T. W., Wan, T. K., Chan, C. L., Wu, C. H., Woo, P. Y. M., Tseng, C. Z. S., Vodencarevic, A., Menni, C., & Chan, K. H. K. (2023). Machine Learning-Based Prediction of COVID-19 Mortality Using Immunological and Metabolic Biomarkers. *BMC Digital Health*, *1*(1), Article ID: 6. https://doi.org/10.1186/s44247-022-00001-0

Witjaksana, E. C. P., Saedudin, Rd. R., & Widartha, V. P. (2021). Perbandingan Akurasi Algoritma Random Forest dan Algoritma Artificial Neural Network untuk Klasifikasi Penyakit Diabetes. *EProceedings of Engineering*, *8*(5), 9773–9781. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15758

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A Review on Over-Sampling Techniques in Classification of Multi-Class Imbalanced Datasets: Insights for Medical Problems. *Frontiers in Digital Health*, *6*, Article ID: 1430245. https://doi.org/10.3389/fdgth.2024.1430245

Zailani, A. U., & Hanun, N. L. (2020). Penerapan Algoritma Klasifikasi Random Forest untuk Penentuan Kelayakan Pemberian Kredit di Koperasi Mitra Sejahtera. *Infotech: Journal of Technology Information*, *6*(1), 7–14. https://doi.org/10.37365/jti.v6i1.61

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 2, MEI, 2025, Pp. 223 – 234
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

LAMPIRAN A

Tabel 9 Dataset

Age	Gender	Family Diabetes	highBP	Physically Active	BMI	Smoking	Alcohol	Sleep	Sound Sleep	Regular Medicine	Junkfood	Strees	BPLevel	Pregancie	Pdiabetes	Urination Frreq
<40	F	No	No	>0.5 hr	21	No	No	7	7	No	Often	Not at all	Normal	0	0	Not much
<40	F	No	No	≥1hr	20	No	No	6	6	No	Very often	Not at all	Normal	0	0	Not much
50-59	M	No	No	≥1hr	27	No	No	7	7	No	Very often	Not at all	Normal	0	0	Not much
40-49	M	Yes	No	None	29	No	No	6	6	No	Occasionally	Not at all	Normal	0	0	Not much
≥60	F	No	No	None	18	No	No	6	6	Yes	Occasionally	Sometimes	Normal	3	0	Quite often

Tabel 10 Hasil Setelah Pengubahan Data

Age	Gender	Family Diabetes	highBP	Physically Active	BMI	Smoking	Alcohol	Sleep	Sound Sleep	Regular Medicine	Junkfood	Strees	BPLevel	Pregancie	Pdiabetes	Urination Frreq
3.0	0.0	0.0	0.0	1.0	21.0	0.0	0.0	7	7	0.0	2.0	1.0	4.0	0.0	0.0	0.0
3.0	0.0	0.0	0.0	3.0	20.0	0.0	0.0	6	6	1.0	3.0	1.0	4.0	0.0	0.0	0.0
1.0	1.0	0.0	0.0	3.0	27.0	0.0	0.0	7	7	0.0	3.0	1.0	4.0	0.0	0.0	0.0
0.0	1.0	1.0	0.0	2.0	29.0	0.0	0.0	6	6	0.0	1.0	1.0	4.0	0.0	0.0	0.0
2.0	0.0	0.0	0.0	2.0	18.0	0.0	0.0	6	6	1.0	1.0	2.0	4.0	3.0	0.0	2.0