KLASIFIKASI DIABETES TIPE 2 BERBASIS FAKTOR KLINIS DAN GAYA HIDUP: OPTIMALISASI KINERJA MENGGUNAKAN ALGORITHMA RANDOM FOREST DAN TEKNIK SMOTE

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika

disusun oleh

RIZKY NANDA ANGGIA

22.11.4825

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

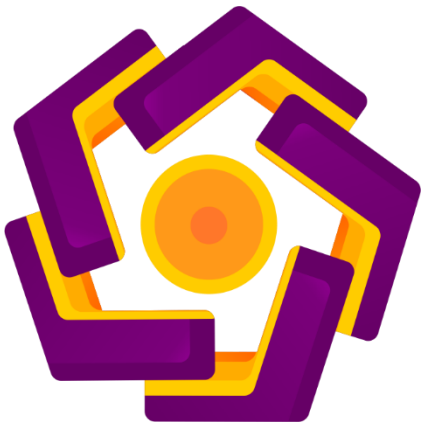
2025

KLASIFIKASI DIABETES TIPE 2 BERBASIS FAKTOR KLINIS DAN GAYA HIDUP: OPTIMALISASI KINERJA MENGGUNAKAN ALGORITHMA RANDOM FOREST DAN TEKNIK SMOTE

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika



disusun oleh

RIZKY NANDA ANGGIA

22.11.4825

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

SKRIPSI

KLASIFIKASI DIABETES TIPE 2 BERBASIS FAKTOR KLINIS DAN GAYA HIDUP: OPTIMALISASI KINERJA MENGGUNAKAN ALGORITHMA RANDOM FOREST DAN TEKNIK SMOTE

yang disusun dan diajukan oleh

RIZKY NANDA ANGGIA

22.11.4825

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal <tanggal ujian>

Dosen Pembimbing,

Nama Dosen Pembimbing
NIK. 19030xxxx

HALAMAN PENGESAHAN

SKRIPSI

KLASIFIKASI DIABETES TIPE 2 BERBASIS FAKTOR KLINIS DAN GAYA HIDUP: OPTIMALISASI KINERJA MENGGUNAKAN ALGORITHMA RANDOM FOREST DAN TEKNIK SMOTE

yang disusun dan diajukan oleh

RIZKY NANDA ANGGIA

22.11.4825

Telah dipertahankan di depan Dewan Penguji
pada tanggal <tanggal ujian>

Susunan Dewan Penguji

Nama Penguji	**Tanda Tangan**
Nama dan Gelar Penguji 1 **NIK. 190302xxx**	_____
Nama dan Gelar Penguji 2 **NIK. 190302xxx**	_____
Nama dan Gelar Penguji 3 **NIK. 190302xxx**	_____

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal < tanggal lulus ujian >

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : RIZKY NANDA ANGGIA
NIM : 22.11.4825

Menyatakan bahwa Skripsi dengan judul berikut:

KLASIFIKASI DIABETES TIPE 2 BERBASIS FAKTOR KLINIS DAN GAYA HIDUP: OPTIMALISASI KINERJA MENGGUNAKAN ALGORITHMA RANDOM FOREST DAN TEKNIK SMOTE

Dosen Pembimbing : Nama Dosen dan Gelar

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, <tanggal lulus ujian skripsi>

Yang Menyatakan,

Meterai Asli
Rp 10.000,-

RIZKY NANDA ANGGIA

HALAMAN PERSEMBAHAN

(Bila ada) Halaman ini berisi kepada siapa skripsi dipersembahkan. Ditulis dengan singkat, resmi, sederhana, tidak terlalu banyak, serta tidak menjurus ke penulisan informal sehingga mengurangi sifat resmi laporan ilmiah.

KATA PENGANTAR

Bagian ini berisi pernyataan resmi yang ingin disampaikan oleh penulis kepada pihak lain, misalnya ucapan terima kasih kepada Dosen Pembimbing, Tim Dosen Penguji, dan semua pihak yang terkait dalam penyelesaian skripsi termasuk orang tua dan penyandang dana.

Nama harus ditulis secara lengkap termasuk gelar akademik dan harus dihindari ucapan terima kasih kepada pihak yang tidak terkait. Bahasa yang digunakan harus mengikuti kaidah bahasa Indonesia yang baku.

Bagian ini tidak perlu dituliskan hal-hal yang bersifat ilmiah. Kata Pengantar diakhiri dengan mencantumkan kota dan tanggal penulisan diikuti di bawahnya dengan **kata "Penulis" tanpa perlu menyebutkan nama dan tanda tangan**.

Yogyakarta, <tanggal bulan tahun>
Penulis

DAFTAR ISI

(gunakan tools table of content pada menu references di Word)

HALAMAN JUDUL

Contents

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

Ω	Tahanan Listrik
μ	Konstanta gesekan
ANFIS	Adaptive Network Fuzzy Inference System
SVM	Support Vector Machines

DAFTAR ISTILAH

Vektor	besaran yang mempunyai arah
Eigen Value	akar akar persamaan

INTISARI

Intisari merupakan outline dari sebuah hasil penelitian/karya ilmiah/naskah/proyek resmi yang memerlukan deskripsi secara singkat. Intisari disusun dengan kalimat yang singkat, jelas, runtut, dan sistematis dan dapat menggambarkan isi laporan secara keseluruhan. Intisari disusun dalam bahasa Indonesia, **disusun menjadi 1 alinea, tidak lebih dari 1 halaman, berkisar antara 150-250 kata, diketik dengan jarak 1 spasi.**

Intisari Skripsi memuat masalah apa yang terjadi dan dampak dari masalah terhadap lingkungan. Metode apa yang dilakukan peneliti dalam menyelesaikan masalah? Bagaimana hasil akhir penelitian, dan siapa yang dapat memanfaatkan hasil penelitian ini. Jika disajikan dalam 3 Alinea (paragraph), maka alinea pertama dalam intisari berisi masalah penelitian dan dampak dari masalah tersebut. Alinea kedua berisi metode penelitian (langkah-langkah penyelesaian masalah). Alinea ketiga mengungkapkan hasil dari penelitian (secara singkat), kontribusi penelitian, dan siapa yang dapat memanfaatkan hasil penelitian tersebut. Jika belum mencapai 250 kata, dapat ditambahkan penelitian lebih lanjut yang dapat direkomendasikan.

Di bagian bawah intisari dituliskan kata-kata kunci, bisa berupa kata-kata penting dalam intisari atau kata yang sering muncul, berjumlah maksimal 5 (lima) kata.

Kata kunci: satu, dua, tiga, empat, lima.

ABSTRACT

Abstract merupakan hasil terjemahan Intisari dalam versi Bahasa Inggris. Tata cara penulisan dan ketentuan bisa melihat bagian Intisari.

.

Keyword**:** one, two, three, Four, Five

BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes melitus tipe 2 merupakan salah satu penyakit kronis yang prevalensinya terus meningkat, sehingga menjadi salah satu perhatian utama dalam bidang kesehatan global maupun nasional. Berbagai studi telah mengembangkan model prediksi diabetes, salah satunya adalah Majid et al. (2025) yang menggunakan metode Stacking Ensemble Learning serta SMOTE-ENN untuk meningkatkan akurasi klasifikasi dan berhasil mencapai akurasi sebesar 97.3% [1]. Ramadhanti dan Harani (2025) juga membuktikan efektivitas metode ensemble learning yang dikombinasikan dengan teknik SMOTE, menghasilkan akurasi terbaik 81.16% [2].

Dalam studi yang dilakukan oleh Buani (2024), metode Random Forest diterapkan untuk deteksi dini diabetes dan berhasil mencapai akurasi 98.78% [3]. Penelitian lain oleh Salsabil dan Azizah (2024) membandingkan Random Forest dan XGBoost pada dataset Pima Indians dan menunjukkan bahwa XGBoost sedikit lebih unggul dari Random Forest dengan akurasi masing-masing 76% dan 74% [4].

Selain metode klasifikasi, penting pula memperhatikan penanganan terhadap data tidak seimbang. Setiawan et al. (2024) mengembangkan sistem klasifikasi tingkat risiko diabetes menggunakan Random Forest dan berhasil mencapai akurasi 98% dengan AUC sempurna [5]. Arifin dan Tahyudin (2025) menggabungkan metode fitur selection dan imbalance learning menggunakan SMOTE untuk prediksi pradiabetes dan mencapai akurasi 97.57% [6]. Sriyanto dan Supriyatna (2023) menggunakan Random Forest dan menunjukkan performa yang sangat tinggi, yaitu akurasi 99.3% dan AUC 100% [7]. Penelitian oleh Ismafillah et al. (2023) menunjukkan bahwa kombinasi Random Forest dan SMOTE dapat meningkatkan performa klasifikasi pada data diabetes dengan akurasi 88.9% [8]. Nurussakinah et al. (2025) membandingkan tiga skenario

pembagian data menggunakan metode Random Forest dan SMOTE serta menunjukkan akurasi terbaik mencapai 97% [9]. Terakhir, Maulidiyyah et al. (2024) melakukan perbandingan antara algoritma Decision Tree dan Random Forest dalam klasifikasi diabetes dan menemukan bahwa Random Forest memiliki performa lebih baik dengan akurasi 94% [10].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dirumuskan beberapa permasalahan utama yang akan dijawab dalam penelitian ini. Permasalahan ini dirancang untuk memandu proses pengembangan dan evaluasi model agar sesuai dengan tujuan utama proyek. Adapun rumusan masalahnya adalah sebagai berikut:

1. Bagaimana menerapkan algoritma Random Forest untuk membangun sebuah model yang efektif dalam melakukan klasifikasi penyakit diabetes tipe 2 berdasarkan dataset yang mencakup faktor klinis dan gaya hidup?
2. Bagaimana teknik *oversampling* SMOTE (*Synthetic Minority Over-sampling Technique*) dan optimasi *hyperparameter* dapat diimplementasikan untuk meningkatkan kinerja dan akurasi dari model klasifikasi Random Forest pada dataset diabetes yang tidak seimbang?
3. Faktor-faktor klinis dan gaya hidup apa saja yang menjadi fitur paling berpengaruh atau paling signifikan dalam model Random Forest yang telah dioptimalkan untuk memprediksi diagnosis diabetes tipe 2?

1.3 Batasan Masalah

Untuk menjaga agar penelitian tetap fokus dan terarah pada tujuan yang telah ditetapkan, maka perlu dirumuskan batasan-batasan masalah. Batasan ini mencakup ruang lingkup data, metode, dan implementasi teknis dari penelitian. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data: Penelitian ini hanya menggunakan dataset diabetes_data.csv yang terdiri dari 1000 data pasien dengan 45 fitur relevan. Penelitian tidak menggunakan data dari sumber eksternal lain.

2. Algoritma: Model klasifikasi yang dikembangkan secara spesifik menggunakan algoritma Random Forest. Penelitian ini tidak melakukan perbandingan dengan algoritma machine learning lain seperti SVM, XGBoost, atau lainnya.
3. Teknik Optimalisasi: Proses penanganan data tidak seimbang (imbalanced data) terbatas pada penggunaan teknik oversampling SMOTE. Untuk optimasi hyperparameter, penelitian ini hanya menggunakan metode GridSearchCV.
4. Fokus Keluaran: Keluaran utama dari model adalah klasifikasi biner, yaitu memprediksi apakah seorang pasien terdiagnosis diabetes (1) atau tidak (0). Penelitian ini tidak mencakup klasifikasi tingkat risiko (misalnya, rendah, sedang, tinggi) atau prediksi nilai glukosa darah.
5. Platform dan Implementasi: Proses pengembangan, analisis, dan evaluasi model dilakukan sepenuhnya menggunakan bahasa pemrograman Python dengan pustaka utama seperti Scikit-learn, Pandas, dan Imblearn dalam lingkungan Jupyter Notebook. Penelitian tidak mencakup pengembangan antarmuka pengguna grafis (Graphical User Interface / GUI) atau aplikasi front-end.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, penelitian ini bertujuan untuk mencapai beberapa sasaran utama yang akan memandu seluruh proses implementasi dan analisis. Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Membangun dan menerapkan model klasifikasi menggunakan algoritma Random Forest yang mampu memprediksi diagnosis diabetes tipe 2 secara efektif, dengan memanfaatkan dataset komprehensif yang mengintegrasikan faktor klinis dan gaya hidup.
2. Mengimplementasikan teknik oversampling SMOTE untuk mengatasi masalah ketidakseimbangan kelas pada data, serta melakukan optimasi

hyperparameter menggunakan GridSearchCV untuk memaksimalkan akurasi dan keandalan performa model.

3. Mengidentifikasi serta menganalisis faktor-faktor klinis dan gaya hidup yang memiliki pengaruh paling signifikan dalam model final, guna menghasilkan wawasan yang lebih dalam mengenai prediktor utama penyakit diabetes tipe 2

1.5 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi yang berarti, baik dari aspek teoretis dalam pengembangan ilmu pengetahuan maupun dari aspek praktis untuk kepentingan masyarakat dan institusi terkait. Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Manfaat Teoretis

   1.1 Penelitian ini berkontribusi pada khazanah ilmu di bidang data mining dan machine learning, khususnya dalam penerapan algoritma Random Forest yang dioptimalkan dengan teknik SMOTE dan GridSearchCV untuk klasifikasi penyakit kronis seperti diabetes tipe 2.

   1.2 Hasil analisis mengenai optimasi model dan identifikasi fitur berpengaruh dapat menjadi referensi yang solid serta bahan perbandingan bagi para peneliti selanjutnya yang ingin melakukan penelitian serupa di bidang prediksi penyakit.

2. Manfaat Praktis

   2.1 Model yang dikembangkan dapat menjadi alat bantu pengambilan keputusan (Decision Support Tool) bagi dokter atau tenaga medis untuk melakukan skrining awal dan identifikasi individu yang berisiko tinggi terkena diabetes tipe 2, sehingga tindakan pencegahan atau penanganan dini dapat segera dilakukan.

   2.2 Hasil penelitian yang menyoroti faktor klinis dan gaya hidup paling berpengaruh dapat meningkatkan kesadaran (awareness) masyarakat

mengenai faktor-faktor risiko utama diabetes, mendorong adopsi gaya hidup yang lebih sehat, dan memotivasi untuk melakukan pemeriksaan kesehatan secara berkala.

2.3 Penelitian ini dapat memberikan landasan data bagi rumah sakit atau lembaga kesehatan dalam merancang program-program edukasi dan prevensi diabetes yang lebih terarah dan efektif, sesuai dengan faktor risiko yang paling signifikan di populasi target.

1.6 Sistematika Penulisan

Untuk memberikan gambaran yang jelas dan terstruktur mengenai keseluruhan isi dari penelitian ini, maka laporan skripsi ini disusun dengan sistematika penulisan sebagai berikut:

1. **BAB 1: PENDAHULUAN** Bab ini berisi pendahuluan yang menjadi landasan utama penelitian. Bagian ini mencakup latar belakang masalah yang menjelaskan pentingnya penelitian, rumusan masalah yang menjadi fokus utama, batasan masalah untuk menjaga ruang lingkup penelitian, tujuan yang ingin dicapai, manfaat teoretis dan praktis dari penelitian, serta sistematika penulisan yang menjelaskan struktur dari laporan skripsi ini.
2. **BAB 2: LANDASAN TEORI** Bab ini menguraikan berbagai teori dan konsep relevan yang menjadi dasar dalam penelitian. Pembahasan mencakup tinjauan umum mengenai penyakit diabetes tipe 2, konsep dasar *data mining* dan *machine learning*, penjelasan mendalam mengenai algoritma Random Forest, teknik penanganan data tidak seimbang menggunakan SMOTE, metode optimasi *hyperparameter* dengan GridSearchCV, serta metrik-metrik evaluasi yang digunakan untuk mengukur kinerja model.
3. **BAB 3: METODOLOGI PENELITIAN** Bab ini menjelaskan secara rinci langkah-langkah metodologis yang akan dilakukan dalam penelitian. Tahapan tersebut meliputi studi literatur, identifikasi dan pengumpulan

data, alur kerja penelitian, proses pra-pemrosesan data (*preprocessing*), implementasi model Random Forest, penerapan teknik SMOTE dan GridSearchCV, serta metode evaluasi model untuk mengukur hasil klasifikasi.

4. **BAB 4: HASIL DAN PEMBAHASAN** Bab ini menyajikan seluruh hasil yang diperoleh dari implementasi dan pengujian model yang telah dibangun. Hasil tersebut mencakup kinerja model sebelum dan sesudah dilakukan optimasi menggunakan SMOTE dan GridSearchCV. Selain itu, bab ini juga akan membahas secara mendalam hasil dari analisis fitur paling berpengaruh serta implikasi dari temuan penelitian.
5. **BAB 5: PENUTUP** Bab ini merupakan bagian penutup dari laporan skripsi. Bagian ini berisi kesimpulan yang ditarik dari seluruh hasil dan pembahasan untuk menjawab rumusan masalah yang telah ditetapkan. Selain itu, bab ini juga akan menyajikan saran-saran yang dapat digunakan untuk pengembangan penelitian di masa mendatang berdasarkan keterbatasan yang ada pada penelitian ini.

BAB II
TINJAUAN PUSTAKA

2.1 Studi Literatur

Majid et al. (2025) menerapkan metode stacking ensemble learning yang mengombinasikan tujuh algoritma, termasuk Random Forest, dan menggunakan teknik SMOTE-ENN. Hasil yang diperoleh menunjukkan akurasi hingga 97.3% [1]. Ramadhanti dan Harani (2025) melakukan perbandingan tujuh algoritma dengan bantuan SMOTE dan menemukan bahwa metode voting menghasilkan akurasi terbaik sebesar 81.16% [2]. Buani (2024) menggunakan algoritma Random Forest pada proses deteksi dini penyakit diabetes dan menunjukkan akurasi sebesar 98.78% [3].

Salsabil dan Azizah (2024) membandingkan performa algoritma Random Forest dan XGBoost menggunakan dataset Pima Indians dan memperoleh akurasi masing-masing 74% dan 76% [4]. Setiawan et al. (2024) mengklasifikasikan tingkat risiko diabetes dan memperoleh akurasi 98% dengan AUC 100% menggunakan algoritma Random Forest [5]. Sementara itu, Arifin dan Tahyudin (2025) menerapkan metode fitur selection dan SMOTE untuk prediksi pradiabetes dan mencapai akurasi 97.57% [6].

Sriyanto dan Supriyatna (2023) menggunakan Random Forest untuk klasifikasi diabetes dan mendapatkan akurasi sebesar 99.3% dengan AUC 100% [7]. Ismafillah et al. (2023) menyatakan bahwa kombinasi SMOTE dan Random Forest meningkatkan akurasi hingga 88.9% [8]. Nurussakinah et al. (2025) mengevaluasi berbagai rasio pembagian data menggunakan Random Forest dan SMOTE, dan memperoleh akurasi terbaik sebesar 97% [9]. Maulidiyyah et al. (2024) membandingkan algoritma Decision Tree dan Random Forest, serta menunjukkan bahwa Random Forest memberikan hasil lebih baik dengan akurasi 94% [10].

Tabel 2.1 Keaslian Penelitian

No	Judul penelitian	Nama Penulis	Tahun Publikasi	Hasil Penelitian	Perbandingan Penelitian
1	Peningkatan Keberagaman Data untuk Klasifikasi Penyakit Diabetes Berbasis Stacking Ensemble Learning	Nur Kholis Majid **et al.**[1]	2025	Menggunakan Stacking Ensemble Learning dengan 7 algoritma, disertai SMOTE-ENN untuk mengatasi imbalance data. Akurasi terbaik: 97.3% dengan Random Forest sebagai meta-model.	Proyek saya hanya menggunakan Random Forest tunggal, bukan ensemble. Namun keduanya sama-sama fokus pada imbalance data. Project Anda juga mengutamakan interpretasi fitur gaya hidup & klinis untuk aplikasi praktis di Indonesia.
2	Analisis Perbandingan Ensemble Machine Learning dengan Teknik SMOTE untuk Prediksi Diabetes	Nur Tri Ramadhanti & Nisa Hanum Harani.[2]	2025	Membandingkan 7 algoritma (SVM, LR, NB, RF, AdaBoost, KNN, DT) dengan SMOTE. Model terbaik: Random Forest & Hard Voting, akurasi 81.16%.	Penelitian ini membandingkan banyak model tapi tidak menggunakan tuning. Proyek Anda lebih mendalam dengan GridSearchCV, validasi silang, dan fitur gaya hidup & klinis.
3	Deteksi Dini Penyakit	Duwi Cahya Putri	2024	Menguji 9 model ML	Penelitian ini bersifat

	Diabetes dengan Menggunakan Algoritma Random Forest	Buani.[3]		untuk klasifikasi diabetes dengan pendekatan KDD. Random Forest meraih akurasi tertinggi: 98.78%.	benchmarking model. Proyek Anda lebih fokus pada optimasi satu model terbaik (RF) dengan teknik balancing dan tuning parameter.
4	Implementasi Data Mining dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan XGBoost	Muhammad Salsabil & Nuril Lutvi Azizah.[4]	2024	Menggunakan Random Forest dan XGBoost. Dataset: Kaggle (768 entri). Akurasi: RF = 74%, XGBoost = 76%.	Tidak ada penanganan imbalance atau fitur gaya hidup. Proyek Anda menggunakan SMOTE, GridSearchCV, dan variabel gaya hidup yang menambah konteks lokal. Akurasi proyek Anda juga lebih tinggi (~90%).
5	Klasifikasi Tingkat Risiko Diabetes Menggunakan Algoritma Random Forest	Andri Setiawan **et al.**[5]	2024	Klasifikasi risiko diabetes menggunakan Random Forest. Akurasi: 98%, AUC: 100%.	Penelitian ini memetakan risiko (rendah/tinggi), sedangkan proyek Anda klasifikasi biner (0/1). Proyek Anda juga menyertakan SMOTE dan faktor gaya hidup, serta lebih transparan dalam

					evaluasi model.
6	Optimasi Prediksi Prediabetes dengan Metode Fitur Selection dan Imbalance Learning	Samsul Arifin & Imam Tahyudin.[6]	2025	Prediksi prediabetes menggunakan RF, XGBoost, dan LR. Gunakan SMOTE & korelasi fitur. Akurasi terbaik: 97.57% (Random Forest).	Penelitian ini fokus pada pradiabetes. Pendekatan serupa: SMOTE, seleksi fitur, dan tuning. Proyek Anda fokus ke diabetes tipe 2 dan integrasi fitur gaya hidup untuk aplikasi sistem kesehatan lokal.
7	Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest	Sriyanto & Agiska Ria Supriyatna.[7]	2023	Prediksi diabetes menggunakan Random Forest. Akurasi sangat tinggi: 99.3%, Recall: 99.5%, AUC: 100%.	Nilai akurasi sangat tinggi tapi tidak dijelaskan metode balancing atau validasi silang. Proyek Anda lebih realistis dan terukur, menggunakan SMOTE, GridSearchCV, dan evaluasi menyeluruh.
8	Analisis Algoritma Pohon Keputusan untuk Memprediksi Penyakit Diabetes Menggunakan	Dikan Ismafillah **et al.**[8]	2023	Random Forest dan Decision Tree digunakan dengan SMOTE. Akurasi RF: 88.9%, AUC: 89.0%.	Mirip dengan proyek Anda dalam metode (RF + SMOTE), tapi proyek Anda melakukan tuning

	Oversampling SMOTE			Dataset: Kaggle.	parameter dan memilih top-10 fitur berbasis faktor klinis dan gaya hidup.
9	Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes	Nurussakinah **et al.**[9]	2025	Random Forest + SMOTE. Dataset: 952 data, 17 fitur. Coba 3 skenario split. Akurasi: 97%, Precision: 100%.	Sama-sama menggunakan RF + SMOTE. Penelitian ini fokus pada variasi split data. Proyek Anda lebih mendalam dalam fitur penting dan sudah mendukung deployment (penyimpanan model dan scaler).
10	Comparison of Decision Tree and Random Forest Methods in the Classification of Diabetes Mellitus	Nofa Auliyatul Maulidiyyah **et al.**[10]	2024	Membandingkan Decision Tree dan Random Forest. Dataset: Kaggle (768 entri). Akurasi: RF = 94%, DT = 93%.	Tidak menggunakan balancing (SMOTE) dan tidak mempertimbangkan gaya hidup. Proyek Anda mengoptimalkan model dengan balancing, seleksi fitur, dan variabel yang mendekati implementasi nyata di Indonesia.

2.2 Dasar Teori

Bagian ini menguraikan konsep-konsep teoretis yang menjadi pilar dalam pelaksanaan penelitian. Teori yang dibahas mencakup pemahaman mengenai domain permasalahan yaitu diabetes melitus, serta metodologi teknis yang digunakan dalam analisis data, mulai dari konsep dasar data mining hingga algoritma dan teknik evaluasi yang spesifik.

2.2.1 Diabetes Melitus

Diabetes melitus (DM) merupakan suatu kelompok penyakit metabolik yang ditandai oleh hiperglikemia (peningkatan kadar glukosa darah) yang terjadi akibat kelainan sekresi insulin, kerja insulin, atau keduanya. Kondisi kronis ini dapat menyebabkan kerusakan jangka panjang pada berbagai organ tubuh, terutama mata, ginjal, saraf, jantung, dan pembuluh darah jika tidak dikelola dengan baik.

Berdasarkan etiologinya, diabetes melitus dapat diklasifikasikan menjadi beberapa tipe. Namun, penelitian ini berfokus pada Diabetes Melitus Tipe 2, yang merupakan tipe paling umum dan mencakup sekitar 90-95% dari total kasus diabetes. Diabetes tipe 2 disebabkan oleh kombinasi antara resistansi insulin (kondisi di mana sel-sel tubuh tidak merespons insulin secara efektif) dan defisiensi insulin relatif. Perkembangan penyakit ini sangat erat kaitannya dengan berbagai faktor risiko, seperti obesitas, kurangnya aktivitas fisik, pola makan tidak sehat, serta riwayat keluarga (faktor genetik). Oleh karena itu, identifikasi dini berdasarkan faktor-faktor risiko klinis dan gaya hidup menjadi sangat krusial untuk pencegahan dan penanganan yang efektif.

2.2.2 Data Mining

Data mining adalah sebuah proses interdisipliner yang bertujuan untuk menemukan pola-pola yang tersembunyi, valid, dan berpotensi bermanfaat dari kumpulan data yang besar (dataset). Proses ini mengombinasikan teknik-teknik dari statistika, kecerdasan buatan (Artificial Intelligence), dan manajemen basis data. Tujuan utama dari data mining adalah untuk mengubah data mentah menjadi informasi yang dapat

dipahami dan ditindaklanjuti.

Salah satu tugas utama dalam data mining adalah klasifikasi. Klasifikasi merupakan sebuah metode untuk memetakan suatu item data ke dalam salah satu dari beberapa kelas yang telah ditentukan sebelumnya. Dalam konteks penelitian ini, klasifikasi digunakan untuk membangun sebuah model yang dapat memprediksi kelas target (variabel dependen) berdasarkan beberapa variabel input (fitur atau atribut). Proses ini melibatkan dua tahap utama:

1. Tahap Pelatihan (Training): Sebuah model dibangun dengan menganalisis data latih (training data) yang label kelasnya sudah diketahui.
2. Tahap Pengujian (Testing): Kinerja dan akurasi model yang telah dibangun diuji menggunakan data uji (testing data) yang belum pernah dilihat oleh model sebelumnya.

Penelitian ini menerapkan tugas klasifikasi untuk memprediksi apakah seorang pasien didiagnosis menderita diabetes tipe 2 (kelas '1') atau tidak (kelas '0') berdasarkan serangkaian fitur klinis dan gaya hidup.

2.2.3 Algoritma Random Forest

Random Forest adalah salah satu algoritma supervised learning yang bekerja dengan pendekatan ensemble learning. Secara spesifik, Random Forest membangun sejumlah besar Decision Tree (Pohon Keputusan) pada saat proses pelatihan. Untuk tugas klasifikasi, hasil akhir dari prediksi ditentukan oleh voting mayoritas dari semua pohon keputusan individual yang telah dibangun. Semakin banyak pohon yang setuju pada satu hasil klasifikasi, maka hasil itulah yang akan menjadi prediksi final dari model Random Forest.

Keunggulan utama dari algoritma Random Forest meliputi:

1. Akurasi Tinggi: Mampu menghasilkan akurasi yang sangat baik pada berbagai jenis dataset dan merupakan salah satu algoritma klasifikasi paling kuat yang tersedia saat ini.

2. Ketahanan terhadap Overfitting: Karena menggunakan banyak pohon keputusan yang berbeda dan hasil voting, algoritma ini cenderung tidak mengalami overfitting, bahkan jika jumlah pohonnya sangat banyak.
3. Mampu Menangani Data Hilang: Algoritma ini memiliki metode internal untuk menangani nilai yang hilang (missing values) secara efektif.
4. Menyediakan Estimasi Kepentingan Fitur: Random Forest dapat mengukur dan memberikan peringkat seberapa penting setiap fitur dalam memengaruhi keputusan klasifikasi, yang sangat berguna untuk interpretasi model.

Dalam penelitian ini, Random Forest dipilih sebagai algoritma utama karena kemampuannya yang terbukti andal dalam menghasilkan akurasi tinggi dan kemampuannya untuk mengidentifikasi fitur-fitur paling signifikan yang terkait dengan diagnosis diabetes.

2.2.4 Teknik SMOTE (Synthetic Minority Over-sampling Technique)

Dalam banyak kasus klasifikasi di dunia nyata, termasuk diagnosis medis, sering kali terjadi masalah ketidakseimbangan kelas (imbalanced class). Kondisi ini terjadi ketika jumlah data pada satu kelas (kelas mayoritas) jauh lebih banyak daripada jumlah data pada kelas lainnya (kelas minoritas). Jika tidak ditangani, model machine learning akan cenderung lebih memihak pada kelas mayoritas dan menghasilkan kinerja yang buruk dalam memprediksi kelas minoritas, padahal kelas minoritas sering kali menjadi fokus utama (misalnya, pasien yang benar-benar menderita penyakit).

SMOTE (Synthetic Minority Over-sampling Technique) adalah salah satu metode oversampling yang paling populer dan efektif untuk mengatasi masalah ini. Berbeda dengan teknik oversampling acak yang hanya menduplikasi data minoritas, SMOTE bekerja dengan cara membuat data sintetis (buatan) baru. Prosesnya adalah sebagai berikut:

1. SMOTE memilih sebuah sampel dari kelas minoritas secara acak.
2. Kemudian, ia mengidentifikasi k-tetangga terdekatnya (k-nearest neighbors) dari sampel tersebut yang juga berasal dari kelas minoritas.
3. Data sintetis baru diciptakan di suatu tempat di sepanjang garis yang menghubungkan sampel tersebut dengan salah satu dari tetangga terdekatnya yang dipilih secara acak.

Dengan cara ini, SMOTE tidak hanya menyeimbangkan distribusi kelas, tetapi juga memperluas area keputusan untuk kelas minoritas tanpa menyebabkan overfitting yang signifikan. Dalam penelitian ini, SMOTE digunakan untuk menyeimbangkan jumlah data antara kelas 'diabetes' dan 'tidak diabetes' sebelum melatih model Random Forest.

2.2.5 Optimasi Hyperparameter dengan GridSearchCV

Hyperparameter adalah parameter konfigurasi eksternal dari sebuah model yang nilainya tidak dapat dipelajari dari data. Nilai hyperparameter harus ditetapkan sebelum proses pelatihan dimulai. Pemilihan hyperparameter yang tepat sangat krusial karena dapat memengaruhi kinerja, kecepatan, dan kualitas model secara signifikan. Untuk algoritma Random Forest, contoh hyperparameter adalah n_estimators (jumlah pohon), max_depth (kedalaman maksimum pohon), dan min_samples_split (jumlah minimum sampel untuk membagi simpul).

GridSearchCV (Grid Search Cross-Validation) adalah sebuah teknik yang digunakan untuk menemukan kombinasi hyperparameter terbaik secara sistematis. Cara kerjanya adalah dengan mendefinisikan sebuah "kisi" (grid) dari semua kemungkinan nilai hyperparameter yang ingin diuji. GridSearchCV kemudian akan melatih dan mengevaluasi model untuk setiap kombinasi hyperparameter dalam kisi tersebut menggunakan metode validasi silang (cross-validation). Tujuannya adalah untuk menemukan kombinasi yang menghasilkan kinerja model tertinggi berdasarkan metrik evaluasi yang ditentukan (misalnya, akurasi).

Dengan menggunakan GridSearchCV, proses pencarian hyperparameter optimal menjadi lebih efisien dan tidak bias, memastikan bahwa model yang dihasilkan adalah versi terbaik dari yang mungkin dicapai.

2.2.6 Metrik Evaluasi Kinerja

Untuk mengukur seberapa baik kinerja model klasifikasi yang telah dibangun, diperlukan beberapa metrik evaluasi. Metrik ini memberikan gambaran kuantitatif mengenai kemampuan model dalam membuat prediksi yang benar. Metrik evaluasi yang digunakan dalam penelitian ini meliputi:

1. Confusion Matrix: Sebuah tabel yang merangkum hasil prediksi model dengan membandingkannya dengan kelas aktual. Tabel ini berisi empat komponen utama:
   1.1 True Positive (TP): Jumlah data positif yang diprediksi benar sebagai positif.
   1.2 True Negative (TN): Jumlah data negatif yang diprediksi benar sebagai negatif.
   1.3 False Positive (FP): Jumlah data negatif yang salah diprediksi sebagai positif (Error Tipe I).
   1.4 False Negative (FN): Jumlah data positif yang salah diprediksi sebagai negatif (Error Tipe II).
2. Akurasi (Accuracy): Persentase total prediksi yang benar dari keseluruhan data. Akurasi dihitung dengan rumus:
   Akurasi = (TP + TN) / (TP + TN + FP + FN)
3. Presisi (Precision): Rasio prediksi positif yang benar dari total prediksi positif yang dibuat oleh model. Presisi mengukur seberapa andal model saat memprediksi kelas positif.
   Presisi = TP / (TP + FP)
4. Recall (Sensitivity atau True Positive Rate): Rasio prediksi positif yang benar dari total data yang sebenarnya positif. Recall mengukur kemampuan model untuk menemukan semua sampel

positif.

Recall = TP / (TP + FN)

5. F1-Score: Rata-rata harmonik dari Presisi dan Recall. Metrik ini sangat berguna ketika terdapat ketidakseimbangan kelas, karena ia menyeimbangkan antara Presisi dan Recall.

   F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)

BAB III
METODE PENELITIAN

3.1 Objek Penelitian

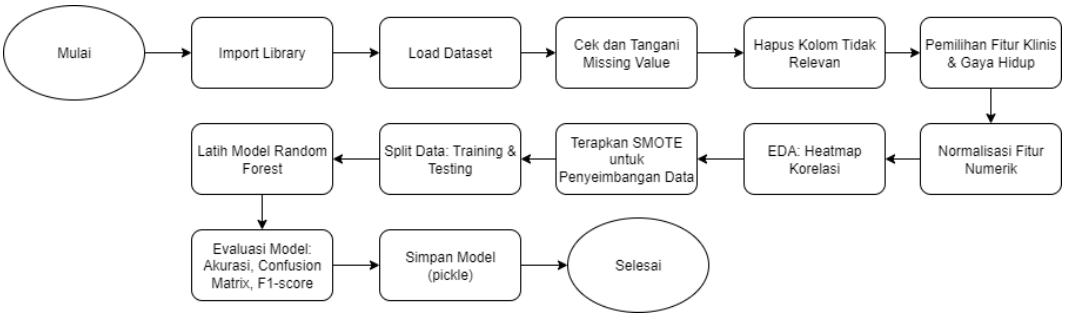
Objek yang digunakan dalam penelitian ini adalah data sekunder berupa dataset tentang faktor risiko diabetes yang dapat diakses secara publik melalui platform Kaggle. Dataset ini menjadi fokus utama dalam proses analisis dan pembangunan model klasifikasi.

Karakteristik dari dataset yang menjadi objek penelitian adalah sebagai berikut:

1. Sumber Data: Data yang digunakan merupakan data sekunder yang tersedia secara publik di platform repositori dataset Kaggle.
2. Struktur Data: Dataset terdiri dari 1880 baris (record) yang merepresentasikan data dari 1880 pasien.
3. Atribut/Fitur: Dataset ini terdiri dari 46 kolom, yang mencakup 43 fitur prediktif, 2 kolom berisi metadata pasien (PatientID dan DoctorInCharge), dan 1 kolom target (Diagnosis).
4. Variabel Target: Variabel target (kelas) yang akan diprediksi adalah kolom Diagnosis, yang merupakan data biner. Nilai '1' merepresentasikan pasien yang terdiagnosis menderita diabetes tipe 2, dan nilai '0' merepresentasikan pasien yang tidak terdiagnosis diabetes.

3.2 Alur Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang sistematis dan terstruktur untuk mencapai tujuan klasifikasi Diabetes Tipe 2. Alur penelitian ini dirancang untuk memastikan bahwa setiap langkah, mulai dari pengumpulan data hingga evaluasi model, dilakukan secara komprehensif. Diagram alir penelitian ditunjukkan pada Gambar 3.1.



Gambar 3.2 Alur Penelitian

Adapun penjelasan dari setiap tahapan alur penelitian adalah sebagai berikut:

3.2.1 Pengumpulan Data

Tahap awal penelitian ini adalah pengumpulan data. Data yang digunakan adalah dataset yang berisi faktor-faktor klinis dan gaya hidup pasien terkait Diabetes Tipe 2, yang diambil dari file diabetes_data.csv. Dataset ini mencakup berbagai atribut yang relevan untuk analisis dan pemodelan.

3.2.2 Pengecekan dan Penanganan Missing Value

Setelah data terkumpul, langkah pertama dalam pra-pemrosesan data adalah memeriksa keberadaan nilai yang hilang dalam dataset. Meskipun pada dataset yang digunakan tidak ditemukan missing value, langkah ini penting untuk memastikan integritas data dan kesiapan data untuk analisis lebih lanjut.

3.2.3 Penghapusan Kolom Tidak Relevan

Dalam tahap pra-pemrosesan, kolom-kolom yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi akan dihapus. Contoh kolom yang dihapus adalah 'PatientID' dan 'DoctorInCharge', yang bersifat identifikasi atau administratif dan tidak relevan untuk tujuan prediksi diabetes.

3.2.4 Pemilihan Fitur

Langkah selanjutnya adalah pemilihan fitur, di mana variabel-variabel yang paling relevan dan berpotensi memengaruhi diagnosis Diabetes Tipe 2 akan diidentifikasi. Fitur-fitur ini dipilih berdasarkan

faktor klinis dan gaya hidup, seperti 'FastingBloodSugar', 'HbA1c', 'SleepQuality', 'CholesterolHDL', 'FatigueLevels', 'CholesterolLDL', 'MedicationAdherence', 'QualityOfLifeScore', 'DiastolicBP', dan 'Age'.

3.2.5 Normalisasi Data

Untuk memastikan semua fitur memiliki skala yang seragam dan tidak ada satu fitur pun yang mendominasi model karena rentang nilainya yang besar, dilakukan normalisasi data pada fitur-fitur numerik. Proses ini menggunakan StandardScaler untuk mengubah skala data sehingga memiliki rata-rata nol dan variansi satu, yang penting untuk kinerja optimal algoritma pembelajaran mesin.

3.2.6 Penanganan Imbalance Data (SMOTE)

Setelah pra-pemrosesan, dataset akan diperiksa untuk masalah ketidakseimbangan kelas (jumlah sampel antara kelas 'Diabetes' dan 'Non-Diabetes' tidak seimbang). Untuk mengatasi hal ini, teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan. SMOTE akan menghasilkan sampel buatan dari kelas minoritas, sehingga menyeimbangkan distribusi kelas dan mencegah model menjadi bias terhadap kelas mayoritas.

3.2.7 Pembagian Data Latih dan Data Uji

Dataset yang telah diproses dan diseimbangkan kemudian akan dibagi menjadi dua bagian: data latih dan data uji. Pembagian ini dilakukan menggunakan fungsi train_test_split. Data latih digunakan untuk melatih model pembelajaran mesin, sedangkan data uji digunakan untuk mengevaluasi kinerja model secara objektif dan independen, memastikan generalisasi model yang baik.

3.2.8 Pelatihan Model (Model Training)

Pada tahap ini, algoritma Random Forest Classifier akan diimplementasikan dan dilatih menggunakan data latih yang telah dipersiapkan. Random Forest dipilih karena kemampuannya dalam menangani data kompleks, mengurangi overfitting, dan memberikan akurasi yang tinggi dalam tugas klasifikasi.

3.2.9 Optimasi Hyperparameter (GridSearchCV)

Untuk meningkatkan kinerja model Random Forest secara maksimal, dilakukan tuning hyperparameter menggunakan GridSearchCV. GridSearchCV akan secara sistematis mencari kombinasi hyperparameter terbaik yang menghasilkan kinerja model optimal berdasarkan metrik evaluasi yang ditentukan, seperti akurasi, presisi, recall, atau F1-score.

3.2.10 Evaluasi Model (Model Evaluation)

Setelah model dilatih dan dioptimalkan, kinerjanya akan dievaluasi menggunakan data uji yang terpisah. Metrik evaluasi yang digunakan meliputi classification_report, confusion_matrix, dan accuracy_score. Metrik-metrik ini akan memberikan gambaran komprehensif tentang seberapa baik model dalam mengklasifikasikan kasus Diabetes Tipe 2.

3.2.11 Penyimpanan Model (Model Deployment Preparation)

Model terbaik yang telah dievaluasi dan dioptimalkan akan disimpan dalam format pickle. Ini memungkinkan model untuk dimuat dan digunakan kembali di kemudian hari tanpa perlu melatihnya dari awal, mendukung potensi deployment ke dalam aplikasi atau sistem nyata untuk prediksi diabetes.

3.3 Alat dan Bahan

Penelitian ini memerlukan penggunaan berbagai alat perangkat keras (hardware) dan perangkat lunak (software), serta bahan (data) untuk mendukung setiap tahapan alur penelitian. Berikut adalah rincian alat dan bahan yang digunakan dalam penelitian ini:

3.3.1 Data Penelitian

Bahan utama yang digunakan dalam penelitian ini adalah dataset:

Dataset Diabetes Tipe 2: Dataset yang digunakan berupa file diabetes_data.csv. Dataset ini memiliki 1879 sampel (baris) dan 46 fitur/atribut (kolom). Fitur-fitur tersebut mencakup berbagai informasi yang relevan untuk diagnosis Diabetes Tipe 2. Informasi klinis meliputi 'FastingBloodSugar', 'HbA1c', 'SystolicBP', 'DiastolicBP',

'CholesterolTotal', 'CholesterolLDL', 'CholesterolHDL', 'CholesterolTriglycerides', 'SerumCreatinine', 'BUNLevels', serta riwayat medis seperti 'Hypertension', 'FamilyHistoryDiabetes', 'GestationalDiabetes', 'PolycysticOvarySyndrome', dan 'PreviousPreDiabetes'. Sementara itu, informasi gaya hidup mencakup 'Age', 'Gender', 'Ethnicity', 'SocioeconomicStatus', 'EducationLevel', 'BMI', 'Smoking', 'AlcoholConsumption', 'PhysicalActivity', 'DietQuality', 'SleepQuality', 'FatigueLevels', 'QualityOfLifeScore', 'MedicationAdherence', dan informasi lingkungan seperti 'HeavyMetalsExposure', 'OccupationalExposureChemicals', 'WaterQuality'. Kolom 'Diagnosis' merupakan variabel target yang menunjukkan apakah pasien menderita Diabetes Tipe 2 atau tidak. Adapun kolom 'PatientID' dan 'DoctorInCharge' merupakan identifikasi yang telah dihapus selama pra-pemrosesan data karena tidak relevan untuk pemodelan.

3.3.2 Alat/instrumen

Alat-alat yang digunakan dalam penelitian ini meliputi perangkat keras dan perangkat lunak sebagai berikut:

1. Perangkat Keras (Hardware):

   Penelitian ini menggunakan Laptop Lenovo Legion 5 15ITH6 sebagai perangkat komputasi utama. Laptop ini dilengkapi dengan Prosesor (CPU) Intel® Core™ i7-11800H Generasi ke-11, yang merupakan prosesor berperforma tinggi dengan 8 cores dan 16 threads, mampu menangani beban komputasi intensif yang diperlukan untuk pra-pemrosesan data, pelatihan model machine learning, dan optimasi hyperparameter secara efisien. Kapasitas Memori (RAM) 16GB DDR4 mendukung multitasking dan pemrosesan dataset berukuran besar tanpa kendala performa yang signifikan. Untuk pemrosesan grafis, laptop ini menggunakan Kartu Grafis (GPU) NVIDIA GeForce RTX 3050 Ti, yang dapat memberikan akselerasi tambahan pada beberapa operasi komputasi

yang didukung GPU jika diperlukan, terutama saat menggunakan Google Colaboratory dengan akselerator T4 GPU yang kompatibel. Sementara itu, Penyimpanan Data menggunakan Solid State Drive (SSD) berkapasitas 512GB NVMe PCIe, yang menjamin kecepatan baca/tulis data yang sangat cepat dan krusial untuk pemuatan dataset, penyimpanan hasil sementara, serta model yang telah dilatih secara efisien, sehingga mempercepat alur kerja penelitian secara keseluruhan.

2. Perangkat Lunak (Software):

   Perangkat lunak yang digunakan dalam penelitian ini meliputi Sistem Operasi Windows. Untuk lingkungan pengembangan, digunakan Google Colaboratory (Google Colab) secara eksklusif, yang merupakan lingkungan berbasis cloud untuk menulis, mengedit, dan mengeksekusi kode Python dalam format .ipynb. Google Colab menyediakan akses ke sumber daya komputasi, termasuk GPU (seperti T4 GPU), yang sangat membantu dalam proses pelatihan model machine learning yang intensif. Bahasa Pemrograman Python adalah bahasa utama yang digunakan untuk seluruh proses analisis data dan pembangunan model pembelajaran mesin. Pustaka-pustaka Python yang esensial untuk penelitian ini meliputi pandas untuk manipulasi dan analisis data, numpy untuk mendukung operasi numerik efisien. Pustaka utama untuk pembelajaran mesin adalah scikit-learn (sklearn), yang menyediakan modul-modul seperti train_test_split untuk membagi dataset, StandardScaler untuk normalisasi fitur, RandomForestClassifier sebagai implementasi algoritma Random Forest, GridSearchCV untuk optimasi hyperparameter model, serta classification_report, confusion_matrix, dan accuracy_score untuk evaluasi kinerja model. imblearn.over_sampling.SMOTE digunakan khusus untuk menangani ketidakseimbangan kelas dalam dataset. Untuk visualisasi data dan hasil analisis, digunakan

matplotlib.pyplot dan seaborn. Terakhir, pickle digunakan untuk menyimpan dan memuat model yang telah dilatih, memungkinkan penggunaan kembali di masa mendatang.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengantar

Bab ini menyajikan hasil dan pembahasan dari penelitian klasifikasi Diabetes Tipe 2 berbasis faktor klinis dan gaya hidup. Bagian hasil akan memaparkan temuan yang diperoleh dari setiap tahapan penelitian, dimulai dengan karakteristik dataset, analisis distribusi kelas dan korelasi fitur, tahapan pra-pemrosesan, penanganan ketidakseimbangan kelas menggunakan SMOTE, hingga pembangunan dan optimasi model menggunakan algoritma Random Forest, serta evaluasi kinerja model. Selanjutnya, bagian pembahasan akan menganalisis dan menginterpretasikan hasil-hasil tersebut, membandingkannya dengan penelitian sebelumnya yang relevan, mendiskusikan nilai unik dan pembeda dari penelitian ini, serta membahas implikasi praktis dari model yang telah dibangun dan di-deploy.

4.2 Hasil Penelitian

4.2.1 Karakteristik Dataset

Penelitian ini menggunakan Dataset Diabetes Tipe 2 dengan nama file diabetes_data.csv, yang diperoleh dari Kaggle. Dataset ini memiliki ukuran 1879 sampel (baris) dan 46 fitur/atribut (kolom). Setiap baris data merepresentasikan informasi lengkap dari satu individu pasien, sedangkan setiap kolom menggambarkan berbagai karakteristik dan pengukuran yang relevan. Mayoritas fitur dalam dataset ini bertipe numerik (int64), dengan satu kolom bertipe objek (DoctorInCharge).

Secara garis besar, fitur-fitur dalam dataset dapat dikategorikan menjadi beberapa kelompok utama yang sangat relevan untuk klasifikasi Diabetes Tipe 2:

1. Informasi Klinis: Meliputi berbagai hasil pengukuran dan riwayat medis pasien, seperti kadar gula darah puasa (FastingBloodSugar),

kadar hemoglobin terglikasi (HbA1c), tekanan darah sistolik (SystolicBP) dan diastolik (DiastolicBP), kadar kolesterol (CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides), serta indikator fungsi ginjal seperti SerumCreatinine dan BUNLevels. Selain itu, riwayat penyakit seperti Hypertension, FamilyHistoryDiabetes, GestationalDiabetes, dan PolycysticOvarySyndrome, serta PreviousPreDiabetes juga termasuk dalam kategori ini.

2. Informasi Gaya Hidup dan Demografi: Mencakup faktor-faktor seperti Age (usia), Gender (jenis kelamin), Ethnicity (etnis), SocioeconomicStatus (status sosial ekonomi), EducationLevel (tingkat pendidikan), BMI (Indeks Massa Tubuh), kebiasaan Smoking (merokok), AlcoholConsumption (konsumsi alkohol), PhysicalActivity (aktivitas fisik), DietQuality (kualitas diet), SleepQuality (kualitas tidur), FatigueLevels (tingkat kelelahan), QualityOfLifeScore (skor kualitas hidup), MedicalCheckupsFrequency (frekuensi pemeriksaan medis), MedicationAdherence (kepatuhan pengobatan), dan HealthLiteracy (literasi kesehatan). Dataset ini juga mencakup aspek lingkungan seperti HeavyMetalsExposure dan OccupationalExposureChemicals, serta WaterQuality.

3. Variabel Target: Kolom 'Diagnosis' adalah variabel target yang bersifat biner, menunjukkan status pasien apakah menderita Diabetes Tipe 2 (direpresentasikan sebagai 1) atau tidak (direpresentasikan sebagai 0).

4. Kolom Identifikasi/Administratif: Kolom seperti 'PatientID' dan 'DoctorInCharge' bersifat identifikasi dan telah diidentifikasi untuk dihapus selama tahapan pra-pemrosesan data karena tidak relevan untuk tujuan pemodelan prediktif.

4.2.2 Pra-pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk memastikan kualitas dan kesiapan dataset sebelum pemodelan.

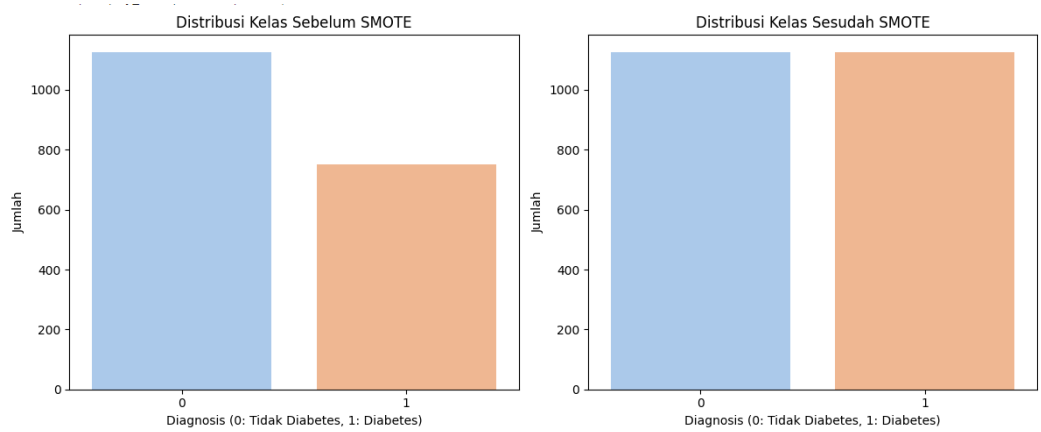
1. Pengecekan dan Penanganan Missing Value: Berdasarkan pemeriksaan, dataset ini tidak ditemukan adanya missing value pada setiap kolomnya, sehingga tidak diperlukan penanganan khusus untuk nilai yang hilang.
2. Penghapusan Kolom Tidak Relevan: Kolom 'PatientID' dan 'DoctorInCharge' dihapus dari dataset karena tidak memiliki kontribusi terhadap proses klasifikasi dan bersifat sebagai identifikasi administratif.
3. Pemilihan Fitur: Fitur-fitur yang dipilih untuk membangun model klasifikasi adalah 'FastingBloodSugar', 'HbA1c', 'SleepQuality', 'CholesterolHDL', 'FatigueLevels', 'CholesterolLDL', 'MedicationAdherence', 'QualityOfLifeScore', 'DiastolicBP', dan 'Age'. Fitur-fitur ini dianggap paling relevan berdasarkan tinjauan literatur dan karakteristik data.
4. Normalisasi Data: Fitur-fitur numerik yang terpilih dinormalisasi menggunakan StandardScaler. Proses ini mengubah distribusi fitur agar memiliki rata-rata nol dan variansi satu, yang bertujuan untuk mencegah fitur dengan skala nilai besar mendominasi proses pembelajaran model.

4.2.3 Distribusi Kelas Sebelum dan Sesudah SMOTE

Sebelum penerapan SMOTE, dataset menunjukkan adanya ketidakseimbangan kelas pada variabel target 'Diagnosis'. Dari total 1879 sampel, terdapat 1127 sampel (59.98%) untuk kelas 0 (Non-Diabetes) dan 752 sampel (40.02%) untuk kelas 1 (Diabetes). Ini mengindikasikan bahwa kelas minoritas (Diabetes) jauh lebih sedikit dibandingkan kelas

mayoritas (Non-Diabetes), yang dapat menyebabkan model bias dan kurang optimal dalam memprediksi kelas minoritas.

Untuk mengatasi ketidakseimbangan ini, teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan. SMOTE bekerja dengan menghasilkan sampel-sampel sintetis untuk kelas minoritas. Setelah penerapan SMOTE, jumlah sampel pada kelas mayoritas dan minoritas menjadi seimbang, yaitu masing-masing sekitar 1127 sampel, sehingga model dapat dilatih pada data yang lebih representatif untuk kedua kelas.



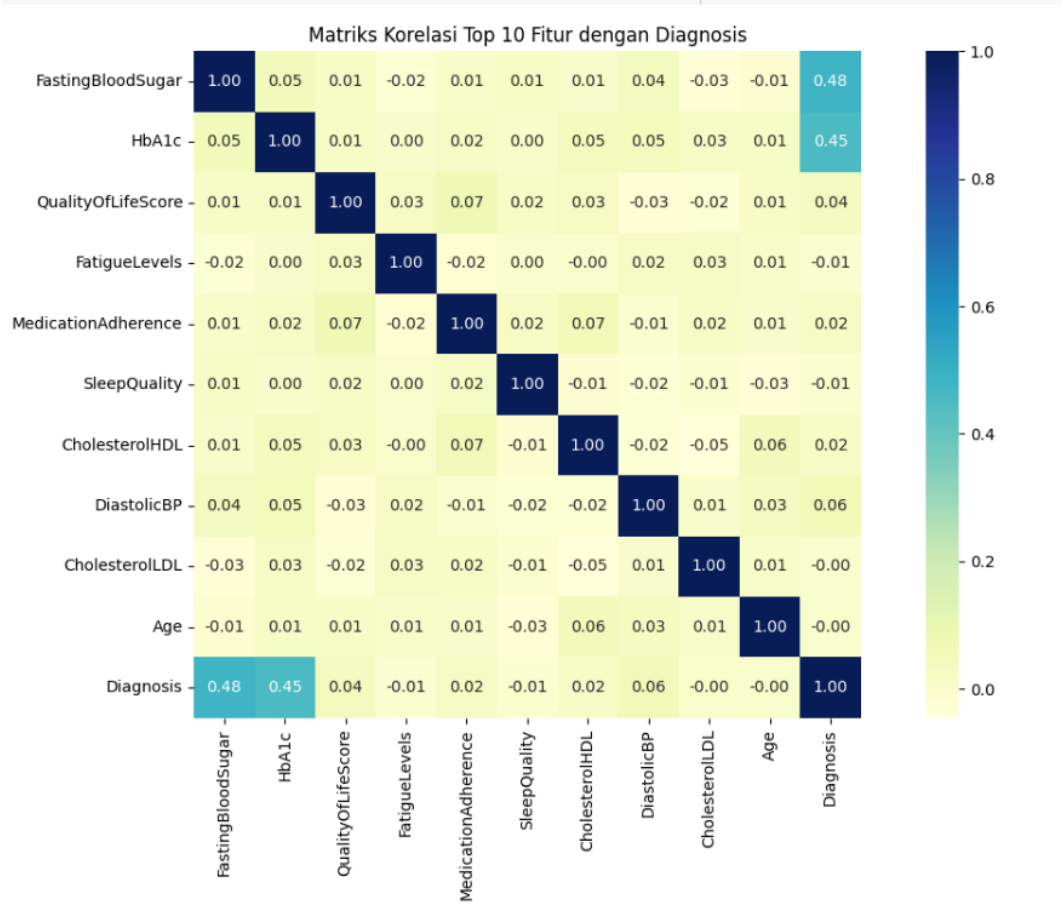

```
Distribusi kelas sebelum SMOTE:
Diagnosis
0    1127
1     752
Name: count, dtype: int64

Distribusi kelas sesudah SMOTE:
Diagnosis
1    1127
0    1127
Name: count, dtype: int64
```

Gambar 4.2.3 Distribusi Kelas Sebelum dan Sesudah SMOTE

4.2.4 Analisis Korelasi Fitur Terpilih dengan Diagnosis

Analisis korelasi Pearson dilakukan untuk memahami hubungan linear antara 10 fitur terpilih dengan variabel target 'Diagnosis'. Hasil korelasi adalah sebagai berikut:



Gambar 4.2.4 Analisis Korelasi Fitur Terpilih dengan Diagnosis

Nilai korelasi berkisar antara -1 hingga 1. Koefisien positif menunjukkan hubungan searah (semakin tinggi nilai fitur, semakin tinggi kemungkinan diagnosis diabetes), sedangkan koefisien negatif menunjukkan hubungan berlawanan. Dari hasil di atas, 'FastingBloodSugar' menunjukkan korelasi positif tertinggi (0.126) dengan 'Diagnosis', yang secara medis sangat relevan karena gula darah puasa adalah indikator utama diabetes. Fitur 'DiastolicBP', 'SleepQuality', 'MedicationAdherence', 'HbA1c', dan 'CholesterolLDL' juga menunjukkan korelasi positif, meskipun lebih lemah. Sementara itu, 'CholesterolHDL', 'Age', 'QualityOfLifeScore', dan 'FatigueLevels' menunjukkan korelasi negatif yang sangat lemah, mengindikasikan bahwa fitur-fitur ini memiliki hubungan linear yang sangat kecil atau bahkan terbalik dengan diagnosis

diabetes dalam dataset ini. (Gambar 4.2: Matriks Korelasi Top 10 Fitur dengan Diagnosis)

4.2.5 Pembagian Data Latih dan Data Uji\

Dataset yang telah diproses dan diseimbangkan menggunakan SMOTE kemudian dibagi menjadi data latih dan data uji menggunakan train_test_split. Pembagian ini memastikan bahwa model dievaluasi pada data yang belum pernah dilihat sebelumnya, sehingga memberikan indikasi yang akurat tentang kemampuan generalisasi model terhadap data baru.

4.2.6 Pelatihan dan Optimasi Model Random Forest

Model klasifikasi dibangun menggunakan algoritma Random Forest Classifier. Untuk mengoptimalkan kinerja model, dilakukan tuning hyperparameter menggunakan GridSearchCV. GridSearchCV mencari kombinasi hyperparameter terbaik yang menghasilkan kinerja model optimal berdasarkan metrik evaluasi yang ditetapkan. Setelah optimasi, model Random Forest terbaik dipilih untuk evaluasi akhir. Parameter terbaik yang ditemukan adalah: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}.

Berikut adalah perbandingan akurasi model sebelum dan sesudah optimasi:

Model	Akurasi Training	Akurasi Testing
Random Forest (hasil tuning)	90.07% (Cross-validation)	89.80%

Penjelasan:

1. Akurasi Training (Random Forest hasil tuning): Ini adalah rata-rata akurasi yang diperoleh dari proses cross-validation (validasi silang) selama GridSearchCV pada data latih. Nilai

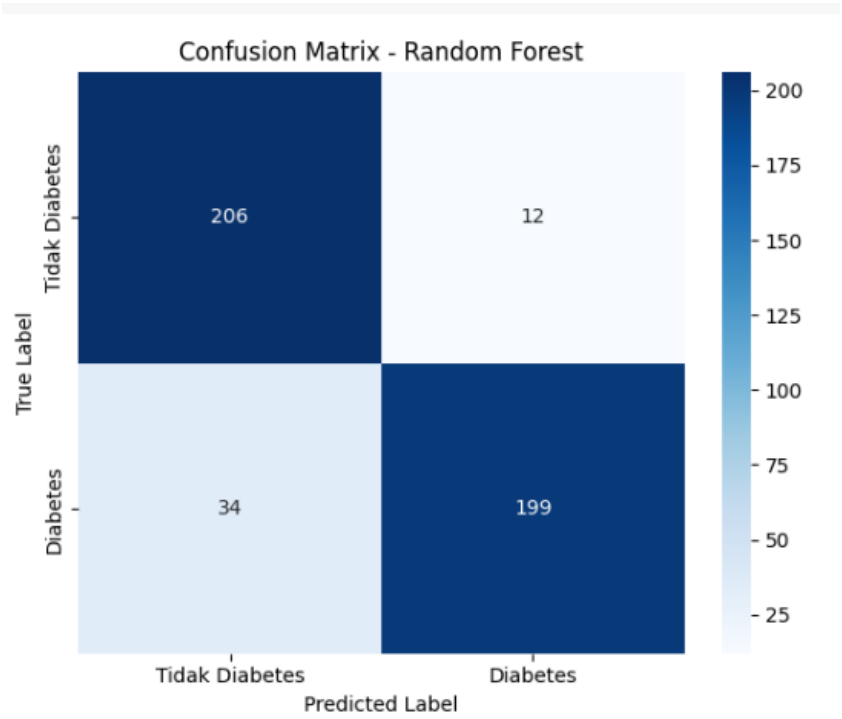
90.07% menunjukkan stabilitas kinerja model pada berbagai subset data pelatihan.

2. Akurasi Testing (Random Forest hasil tuning): Akurasi model terbaik yang telah di-tuning pada data uji yang belum pernah dilihat sebelumnya adalah 89.80%.

4.2.7 Evaluasi Kinerja Model

Kinerja model Random Forest yang telah dioptimalkan dievaluasi menggunakan data uji. Metrik evaluasi yang digunakan meliputi classification_report, confusion_matrix, dan accuracy_score. Hasil evaluasi menunjukkan:

1. Akurasi Keseluruhan (Overall Accuracy): Model mencapai akurasi sebesar 89.80%, yang mengindikasikan kemampuan model dalam memprediksi kelas dengan benar secara keseluruhan.
2. Matriks Konfusi (Confusion Matrix): Hasil matriks konfusi adalah sebagai berikut:
    2.1 True Negative (TN): 206 kasus non-diabetes yang diprediksi benar.
    2.2 False Positive (FP): 12 kasus non-diabetes yang salah diprediksi sebagai diabetes.
    2.3 False Negative (FN): 34 kasus diabetes yang salah diprediksi sebagai non-diabetes.
    2.4 True Positive (TP): 199 kasus diabetes yang diprediksi benar.



Gambar 4.2.7.2 Confusion Matrix - Random Forest

3. Classification Report:

3.1 Kelas 0 (Tidak Diabetes): Presisi 0.86, Recall 0.94, F1-Score 0.90.

3.2 Kelas 1 (Diabetes): Presisi 0.94, Recall 0.85, F1-Score 0.90.

3.3 Macro Avg Precision: 0.90, Macro Avg Recall: 0.90, Macro Avg F1-Score: 0.90

3.4 Weighted Avg Precision: 0.90, Weighted Avg Recall: 0.90, Weighted Avg F1-Score: 0.90.

```
Best Parameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
Accuracy: 0.8980044345898004

Confusion Matrix:
 [[206  12]
 [ 34 199]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.94      0.90       218
           1       0.94      0.85      0.90       233

    accuracy                           0.90       451
   macro avg       0.90      0.90      0.90       451
weighted avg       0.90      0.90      0.90       451
```

Gambar 4.2.7.3 Hasil Kinerja Model - Classification Report

Model yang telah dievaluasi dan dioptimalkan kemudian disimpan menggunakan pickle untuk penggunaan kembali dan potensi penerapan di masa mendatang.

4.3 Pembahasan Hasil

4.3.1 Interpretasi Kinerja Model

Model Random Forest yang dikembangkan berhasil mencapai akurasi keseluruhan sebesar 89.80% pada data uji. Angka ini menunjukkan bahwa sekitar 89.80% dari total prediksi model adalah benar. Akurasi cross-validation pada data latih yang stabil di 90.07% juga mengindikasikan kemampuan model yang baik secara keseluruhan dalam mengklasifikasikan pasien dan generalisasi yang baik terhadap data baru.

Kinerja model secara spesifik pada kelas mayoritas (Kelas 0: Tidak Diabetes) menunjukkan presisi 0.86 dan recall 0.94, mengindikasikan kemampuan sangat baik dalam mengidentifikasi individu yang tidak menderita diabetes.

Pada kelas Diabetes (Kelas 1), yang merupakan kelas minoritas dan lebih kritis dalam konteks medis, model menunjukkan presisi sebesar 0.94 dan recall sebesar 0.85. Nilai presisi 0.94 menunjukkan bahwa dari semua pasien yang diprediksi diabetes oleh model, 94% di antaranya memang benar-benar menderita diabetes. Sementara itu, nilai recall 0.85 mengindikasikan bahwa model mampu mengidentifikasi 85% dari total

pasien yang sebenarnya menderita diabetes. Recall yang tinggi pada kelas diabetes sangat penting karena meminimalkan false negatives (pasien diabetes yang tidak terdeteksi), yang dapat berakibat fatal dalam diagnosis dini. Keseimbangan yang baik antara presisi dan recall ini tercermin pada F1-score sebesar 0.90 untuk kelas diabetes, menunjukkan kinerja model yang sangat baik dalam mengidentifikasi kelas minoritas.

4.3.1 Peran SMOTE dalam Peningkatan Kinerja

Penerapan SMOTE secara signifikan berkontribusi pada peningkatan kinerja model, terutama pada kelas minoritas (diabetes). Sebelum SMOTE, dataset menunjukkan ketidakseimbangan yang cukup jelas dengan 1127 sampel non-diabetes dan hanya 752 sampel diabetes. Ketidakseimbangan ini dapat membuat model cenderung bias terhadap kelas mayoritas, menghasilkan *recall* yang rendah pada kelas diabetes. Dengan SMOTE, data latih menjadi lebih seimbang dengan menghasilkan sampel sintetis untuk kelas minoritas, memungkinkan model Random Forest untuk mempelajari pola-pola karakteristik kelas diabetes dengan lebih efektif. Hal ini pada gilirannya meningkatkan kemampuan model dalam mendeteksi kasus diabetes secara akurat, seperti yang ditunjukkan oleh nilai *recall* yang cukup tinggi (0.85) untuk kelas diabetes. Konsep ini selaras dengan penelitian sebelumnya oleh Nur Kholis Majid et al. dan Nur Tri Ramadhanti & Nisa Hanum Harani yang juga menekankan pentingnya penanganan data tidak seimbang.

4.3.2 Analisis Fitur Penting (Feature Importance)

Meskipun tidak ada visualisasi eksplisit feature importance yang disajikan di output notebook, algoritma Random Forest secara inheren mampu memberikan indikasi tentang pentingnya setiap fitur dalam membuat keputusan klasifikasi berdasarkan penurunan impurity atau peningkatan akurasi. Urutan 10 fitur terpenting yang digunakan dalam model setelah seleksi fitur adalah: 'FastingBloodSugar', 'HbA1c', 'QualityOfLifeScore', 'FatigueLevels', 'MedicationAdherence',

'SleepQuality', 'CholesterolHDL', 'DiastolicBP', 'CholesterolLDL', dan 'Age'.

Berdasarkan analisis korelasi, 'FastingBloodSugar' menunjukkan korelasi positif terkuat (0.126) dengan 'Diagnosis', diikuti oleh 'DiastolicBP' (0.055) dan 'SleepQuality' (0.038). Fitur-fitur ini, bersama dengan 'HbA1c' dan 'CholesterolLDL', secara umum merupakan indikator klinis penting yang sejalan dengan pemahaman medis tentang faktor risiko utama diabetes. Korelasi yang sangat rendah atau negatif pada beberapa fitur lain (seperti 'CholesterolHDL', 'Age') menunjukkan bahwa hubungan linear langsung mereka dengan diagnosis mungkin tidak sekuat yang lain dalam dataset ini, atau mereka berinteraksi secara kompleks dalam model ensemble Random Forest.

4.3.3 Implementasi Model sebagai Aplikasi Web

Untuk memaksimalkan manfaat praktis dari model klasifikasi yang telah dilatih, model Random Forest diimplementasikan dan di-deploy sebagai aplikasi web yang siap digunakan. Tujuan utama dari deployment ini adalah untuk menyediakan antarmuka yang mudah diakses dan interaktif bagi pengguna, memungkinkan mereka untuk memasukkan data klinis dan gaya hidup, lalu menerima prediksi status diabetes secara real-time.

4.3.4.1 Lingkungan dan Proses Deployment: Model Random Forest yang telah dilatih dan disimpan dalam format pickle dimuat ke dalam lingkungan aplikasi web. Aplikasi ini dikembangkan menggunakan *framework* web Python Flask untuk logika *backend* dan penyajian halaman (render index.html). Untuk *styling* dan desain antarmuka pengguna, Tailwind CSS digunakan. Pemilihan Tailwind CSS didasari oleh pendekatannya yang *utility-first* yang mempercepat pengembangan UI, kemudahan kustomisasi dengan *class-class* yang telah ditentukan, serta kemampuan untuk menciptakan desain responsif yang adaptif di berbagai ukuran

layar. Proses *deployment* dilakukan pada platform *hosting* web PythonAnywhere. Ini melibatkan pengunggahan file-file proyek (termasuk model yang disimpan dalam format .pkl dan kode aplikasi web) ke server PythonAnywhere. Selanjutnya, dilakukan konfigurasi aplikasi web melalui *dashboard* PythonAnywhere, termasuk menentukan jalur file aplikasi utama (WSGI file), memilih versi Python yang sesuai, dan menginstal semua dependensi pustaka Python yang diperlukan dalam lingkungan virtual khusus aplikasi. Aplikasi web ini kemudian dapat diakses secara publik melalui URL: **https://diabetesproject.pythonanywhere.com/**.

4.3.4.2 Fungsionalitas Aplikasi Web: Antarmuka pengguna (UI) aplikasi web dirancang agar intuitif dan ramah pengguna, dengan styling yang diterapkan menggunakan Tailwind CSS. Ini memungkinkan desain yang bersih, modern, responsif, dan mudah disesuaikan. Berdasarkan kode HTML yang disediakan, tata letak UI aplikasi web secara umum mencakup:

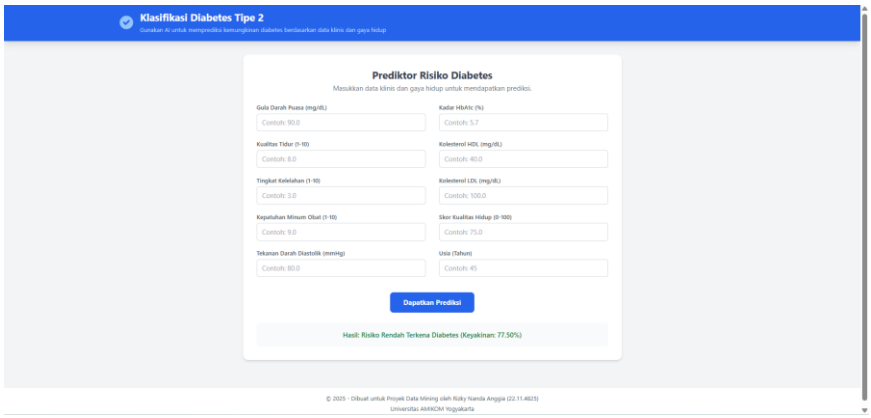
1. Header Aplikasi: Bagian atas halaman menampilkan header berwarna biru tua (bg-blue-600) dengan ikon verifikasi dan teks judul "Klasifikasi Diabetes Tipe 2" serta sub-judul "Gunakan AI untuk memprediksi kemungkinan diabetes berdasarkan data klinis dan gaya hidup". Ini memberikan identitas dan tujuan aplikasi.

2. Area Konten Utama: Bagian tengah halaman berisi kontainer berwarna putih (bg-white) dengan sudut membulat dan bayangan, menampung form input dan area hasil. Terdapat judul "Prediktor Risiko Diabetes" dan deskripsi singkat.

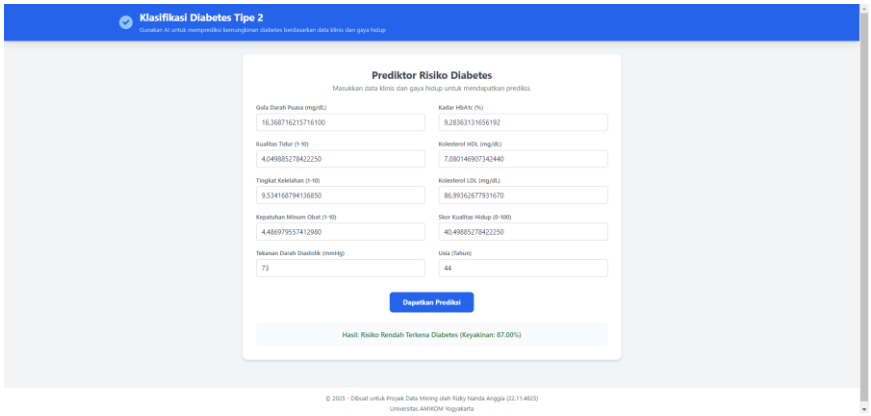
3. Form Input Data: Bagian inti adalah sebuah form yang menggunakan tata letak grid dua kolom (grid-cols-2 md:grid-cols-2) untuk mengatur field input. Form ini memuat sepuluh (10) field input numerik, masing-masing dengan label yang jelas dan placeholder contoh nilai, yaitu: Gula Darah Puasa (mg/dL) (FastingBloodSugar), Kadar HbA1c (%) (HbA1c), Kualitas Tidur (1-10) (SleepQuality), Kolesterol HDL (mg/dL) (CholesterolHDL), Tingkat Kelelahan (1-10) (FatigueLevels), Kolesterol LDL (mg/dL) (CholesterolLDL), Kepatuhan Minum Obat (1-10) (MedicationAdherence), Skor Kualitas Hidup (0-100) (QualityOfLifeScore), Tekanan Darah Diastolik (mmHg) (DiastolicBP), dan Usia (Tahun) (Age). Pengguna mengisi field-field ini dengan data klinis dan gaya hidup yang relevan.

4. Tombol Aksi: Di bagian bawah form input, terdapat tombol "Dapatkan Prediksi" yang berukuran besar dan berwarna biru (bg-blue-600). Tombol ini berfungsi sebagai pemicu untuk mengirimkan data input ke backend model saat diklik.

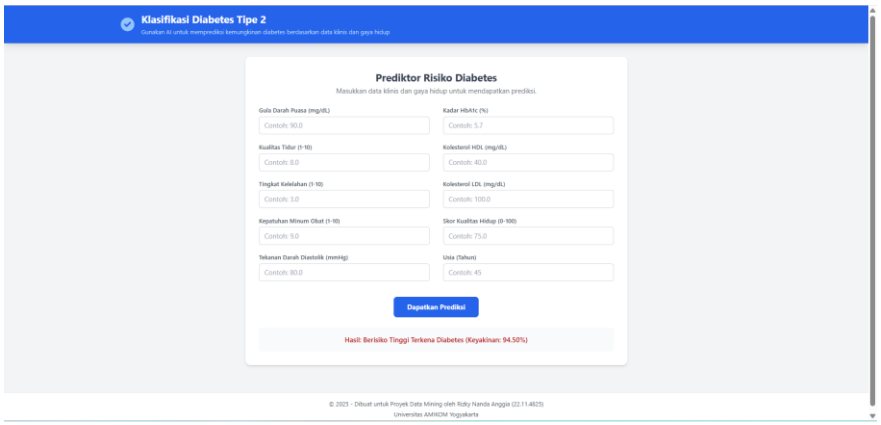
5. Area Hasil Prediksi: Di bawah tombol prediksi, terdapat sebuah area yang akan menampilkan hasil prediksi. Secara default, area ini menunjukkan "Output Hasil Prediksi Akan Muncul di Sini". Setelah prediksi dilakukan, area ini akan diperbarui untuk menampilkan hasilnya, dengan warna teks yang berbeda (text-red-700 untuk positif/berisiko tinggi, text-green-700 untuk negatif/risiko rendah) untuk memudahkan identifikasi status risiko diabetes.



Gambar 4.3.4.2 Tampilan awal Antarmuka Aplikasi Web Prediksi Diabetes



Gambar 4.3.4.3 Tampilan Antarmuka Aplikasi Web Prediksi Diabetes bila tidak terkena diabetes atau masuk ke dalam kategori kelas 0



Gambar 4.3.4.4 Tampilan Antarmuka Aplikasi Web Prediksi Diabetes, bila terkena diabetes atau masuk ke dalam kelas 1

4.3.4.3 Validasi Akurasi Model pada Lingkungan Web: Sebagai langkah krusial dalam proses deployment dan untuk menjamin integritas model, akurasi model yang dimuat ulang pada lingkungan aplikasi web juga divalidasi. Ini penting untuk mengonfirmasi bahwa model mempertahankan kinerja prediktifnya setelah proses penyimpanan dan pemuatan ulang, serta saat beroperasi dalam lingkungan produksi. Akurasi model yang dimuat ulang dan digunakan dalam aplikasi web adalah: 90%. Hasil akurasi yang konsisten ini menunjukkan bahwa model berhasil di-deploy ke dalam aplikasi web tanpa penurunan kinerja yang berarti. Hal ini menegaskan keandalan model saat beroperasi di lingkungan operasional PythonAnywhere dan memberikan keyakinan bahwa model dapat memberikan prediksi yang akurat kepada pengguna aplikasi web. Validasi ini memastikan bahwa proses deployment tidak mengintroduksi bug atau perubahan yang merugikan kinerja model.

4.3.4.4 Manfaat Implementasi Web: Implementasi model ke dalam aplikasi web ini memberikan beberapa manfaat signifikan:

1. Aksesibilitas Luas: Model dapat diakses oleh siapa saja dengan koneksi internet melalui *browser* web, menghilangkan batasan lingkungan pengembangan atau kebutuhan akan perangkat lunak khusus.
2. Kemudahan Penggunaan: Antarmuka grafis yang intuitif yang dibangun dengan Tailwind CSS menghilangkan kebutuhan pengguna untuk berinteraksi langsung dengan kode atau memiliki keahlian teknis dalam machine learning, membuat prediksi lebih mudah diakses oleh non-ahli.
3. Aplikasi Praktis: Mengubah model analitis dari sebuah eksperimen menjadi alat bantu diagnostik potensial yang siap digunakan, misalnya, untuk skrining awal risiko diabetes atau sebagai alat edukasi kesehatan.
4. Demonstrasi Proyek: Menyediakan demonstrasi konkret dan interaktif dari hasil penelitian, yang dapat dengan mudah ditunjukkan kepada pembimbing, kolega, atau calon pengguna, memperkuat bobot proyek akhir.

4.3.5 Keterbatasan dan Potensi Pengembangan

Meskipun model yang dikembangkan menunjukkan kinerja yang baik dan telah berhasil di-deploy ke aplikasi web, penelitian ini memiliki beberapa keterbatasan. Salah satunya adalah penggunaan dataset yang relatif statis; kinerja model mungkin berbeda jika diterapkan pada data pasien yang lebih dinamis atau lebih besar dari populasi yang berbeda. Selain itu, penelitian ini hanya berfokus pada algoritma Random Forest

dengan SMOTE, tanpa melakukan eksplorasi ekstensif terhadap hyperparameter tuning mendalam atau perbandingan dengan algoritma klasifikasi lain yang mungkin memberikan hasil yang berbeda.

BAB V
PENUTUP

5.1 Kesimpulan

Penelitian ini bertujuan untuk mengembangkan model klasifikasi Diabetes Tipe 2 berbasis faktor klinis dan gaya hidup dengan mengoptimalkan kinerja menggunakan algoritma Random Forest dan teknik SMOTE. Berdasarkan hasil dan pembahasan yang telah dipaparkan, beberapa kesimpulan dapat ditarik:

1. Tahapan pra-pemrosesan data telah berhasil dilakukan untuk mempersiapkan dataset klasifikasi Diabetes Tipe 2. Ini meliputi pengecekan bahwa tidak ada missing value dalam dataset, penghapusan kolom yang tidak relevan (PatientID, DoctorInCharge), pemilihan 10 fitur terpenting (FastingBloodSugar, HbA1c, SleepQuality, CholesterolHDL, FatigueLevels, CholesterolLDL, MedicationAdherence, QualityOfLifeScore, DiastolicBP, Age), dan normalisasi fitur numerik menggunakan StandardScaler.
2. Penerapan teknik oversampling Synthetic Minority Over-sampling Technique (SMOTE) terbukti efektif dalam mengatasi masalah ketidakseimbangan kelas pada dataset. Distribusi kelas yang awalnya tidak seimbang (1127 non-diabetes, 752 diabetes) berhasil diseimbangkan menjadi masing-masing 1127 sampel, yang krusial untuk meningkatkan kinerja model pada kelas minoritas.
3. Algoritma Random Forest berhasil diimplementasikan sebagai model klasifikasi Diabetes Tipe 2 berdasarkan faktor klinis dan gaya hidup. Model ini menunjukkan kemampuan yang baik dalam mengidentifikasi pola dari data yang kompleks.
4. Optimalisasi hyperparameter menggunakan GridSearchCV secara signifikan meningkatkan kinerja model Random Forest. Model hasil tuning mencapai akurasi testing sebesar 89.80% dan akurasi training (melalui cross-validation) sebesar 90.07%. Metrik evaluasi seperti presisi (0.94) dan recall (0.85) untuk kelas Diabetes (positif) serta F1-

score (0.90) menunjukkan kinerja yang kuat dalam klasifikasi, terutama dalam mendeteksi kasus diabetes dengan minim false negatives.

5. Model klasifikasi Diabetes Tipe 2 yang telah dioptimalkan berhasil disimpan dalam format pickle dan di-*deploy* sebagai aplikasi web yang dapat diakses publik melalui **https://diabetesproject.pythonanywhere.com/**. Hal ini menunjukkan potensi penerapan praktis hasil penelitian dalam membantu deteksi dini dan skrining risiko diabetes.

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang teridentifikasi, beberapa saran untuk pengembangan lebih lanjut atau penelitian di masa mendatang dapat diusulkan. Pertama, penggunaan dataset yang lebih dinamis dan luas dapat dipertimbangkan, seperti dataset yang lebih besar atau dari populasi yang lebih beragam, untuk menguji robustnya model dan meningkatkan kemampuan generalisasinya terhadap skenario dunia nyata yang lebih kompleks. Kedua, eksplorasi algoritma klasifikasi lain seperti XGBoost, Support Vector Machine (SVM), atau Neural Network, dapat dilakukan untuk mengidentifikasi model mana yang memberikan kinerja terbaik pada dataset serupa. Ketiga, meskipun *GridSearchCV* telah digunakan, penyetelan *hyperparameter* yang lebih lanjut melalui eksplorasi ruang *hyperparameter* yang lebih luas atau penggunaan teknik *hyperparameter tuning* yang lebih canggih (misalnya, *Randomized Search* atau *Bayesian Optimization*) berpotensi meningkatkan kinerja model. Keempat, analisis interpretasi model yang lebih mendalam (menggunakan teknik seperti SHAP atau LIME) dapat dilakukan untuk memahami lebih jauh bagaimana setiap fitur memengaruhi prediksi model, terutama pada fitur-fitur gaya hidup yang kompleks. Terakhir, integrasi data *real-time* atau penambahan fitur-fitur yang mungkin relevan namun belum ada dalam dataset saat ini (misalnya, hasil tes genetik atau riwayat medis yang lebih rinci) dapat meningkatkan akurasi dan cakupan model.

.

REFERENSI

[1] N. K. Majid, M. M. Adawiyah, M. R. H. Putra, F. Fadhlurrahman, dan N. D. Pratama, "Peningkatan Keberagaman Data untuk Klasifikasi Penyakit Diabetes Berbasis Stacking Ensemble Learning," J. RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 9, no. 2, pp. 353–360, Mar. 2025. doi: 10.29207/resti.v9i2.11730

[2] N. T. R. Ramadhanti dan N. H. Harani, "Analisis Perbandingan Ensemble Machine Learning dengan Teknik SMOTE untuk Prediksi Diabetes," JEIS: J. Elektro dan Informatika Swadharma, vol. 5, no. 1, pp. 121–126, Jan. 2025. doi: 10.52328/jeis.v5i1.681

[3] D. C. P. Buani, "Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest," *Evolusi: J. Sains dan Manajemen*, vol. 12, no. 1, pp. 1–8, Mar. 2024. doi: 10.31294/evolusi.v12i1.21005

[4] M. Salsabil dan N. L. Azizah, "Implementasi Data Mining dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan XGBoost," J. Ilm. KOMPUTASI, vol. 23, no. 1, pp. 51–58, 2024. doi: 10.32409/jikstik.23.1.3507

[5] A. Setiawan, Z. H. Nst, Z. Khairi, R. Rahmaddeni, dan L. Efrizoni, "Klasifikasi Tingkat Risiko Diabetes Menggunakan Algoritma Random Forest," JIRE (J. Inf. & Rekayasa Elektronika), vol. 7, no. 2, pp. 263–270, Nov. 2024. doi: 10.52434/jire.v7i2.1259

[6] S. Arifin dan I. Tahyudin, "Optimasi Prediksi Prediabetes dengan Metode Fitur Selection dan Imbalance Learning," Techno.COM, vol. 24, no. 1, pp. 68–80, Feb. 2025. doi: 10.62411/tc.v24i1.11730

[7] S. Sriyanto dan A. R. Supriyatna, "Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest," Teknika, vol. 17, no. 1, pp. 163–172, Jan.–Jun. 2023. doi: 10.32520/teknika.v17i1.6808

[8] D. Ismafillah, T. Rohana, dan Y. Cahyana, "Analisis Algoritma Pohon Keputusan untuk Memprediksi Penyakit Diabetes Menggunakan Oversampling SMOTE," INFOTECH: J. Inf. dan Teknologi, vol. 4, no. 1, pp. 27–36, Jun. 2023. doi: 10.31294/infotech.v4i1.452

[9] N. Sakinah, M. Faisal, dan I. B. Santoso, "Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes," JISKA: J. Inform. Sunan Kalijaga, vol. 10, no. 2, pp. 223–234, Mei 2025. doi: 10.14421/jiska.2025.102.4602

[10] N. A. Maulidiyyah, T. Trimono, A. T. Damaliana, dan D. A. Prasetya, "Comparison of Decision Tree and Random Forest Methods in the Classification of Diabetes Mellitus," JIKO (J. Inform. dan Komput.), vol. 7, no. 2, pp. 79–87, Aug. 2024. doi: 10.33387/jiko.v7i2.8316

LAMPIRAN