TESIS

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SKRINING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

disusun oleh

Nama : **Muh Ikbal Sodikin**

NIM : **22.55.2291**

Konsentrasi : **Business Intelligence**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

TESIS

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SKRINING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

DEVELOPMENT OF A HYBRID ENSEMBLE-BASED DIABETES PREDICTION MODEL WITH SMOTE INTEGRATION FOR SCREENING OPTIMIZATION IN HOSPITAL MANAGEMENT INFORMATION SYSTEMS

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi PJJ S2 INFORMTIKA

disusun oleh

Nama : **Muh Ikbal Sodikin**

NIM : **22.55.2291**

Konsentrasi : **Business Intelligence**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SKRINING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

DEVELOPMENT OF A HYBRID ENSEMBLE-BASED DIABETES PREDICTION MODEL WITH SMOTE INTEGRATION FOR SCREENING OPTIMIZATION IN HOSPITAL MANAGEMENT INFORMATION SYSTEMS

yang disusun dan diajukan oleh

Muh Ikbal Sodikin

22.55.2291

telah disetujui oleh Dosen Pembimbing Tesis

03 Feb 2026

Dosen Pembimbing,

Prof. Dr. Ema Utami, S.Si., M.Kom.

NIK. 190302037

HALAMAN PENGESAHAN

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SKRINING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

DEVELOPMENT OF A HYBRID ENSEMBLE-BASED DIABETES PREDICTION MODEL WITH SMOTE INTEGRATION FOR SCREENING OPTIMIZATION IN HOSPITAL MANAGEMENT INFORMATION SYSTEMS

yang disusun dan diajukan oleh

Muh Ikbal Sodikin

22.55.2291

Telah dipertahankan di depan Dewan Penguji

pada tanggal Tanggal 03 Feb 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Alva Hendi Muhammad, A.Md., S.T., M.Eng., Ph.D.	
NIK. 190302493	
Dhani Ariatmanto, S.Kom., M.Kom., Ph.D.	
NIK. 190302197	
Prof. Dr. Ema Utami, S.Si., M.Kom	
NIK. 190302037	

Tesis ini telah diterima sebagai salah satu persyaratan

untuk memperoleh gelar Magister Komputer

Tanggal 03 Feb 2026

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.

NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : **Muh Ikbal Sodikin**

NIM : **22.55.2291**

Menyatakan bahwa Tesis dengan judul berikut:

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SCREENING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

Dosen Pembimbing : Prof. Dr. Ema Utami, S.Si., M.Kom

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.

1. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
2. 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.
3. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
4. 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 03 Feb 2026

Muh Ikbal Sodikin

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur ke hadirat Allah SWT atas segala rahmat, karunia, serta kemudahan yang telah diberikan, karya ilmiah ini penulis persembahkan kepada: Kedua orang tua tercinta, yang senantiasa memberikan kasih sayang, doa, dukungan moral maupun material, serta menjadi sumber kekuatan dan inspirasi dalam setiap langkah perjalanan pendidikan penulis. Keluarga dan orang-orang terdekat, yang selalu memberikan semangat, motivasi, dan doa sehingga penulis dapat menyelesaikan penelitian ini dengan baik. Dosen pembimbing dan seluruh civitas akademika, yang telah memberikan ilmu, arahan, bimbingan, serta dukungan selama proses penyusunan penelitian ini. Teman-teman seperjuangan, yang telah menjadi bagian dari perjalanan akademik penulis, saling memberikan motivasi, bantuan, dan semangat dalam menyelesaikan tugas akhir ini. Almamater tercinta, yang telah menjadi tempat penulis menimba ilmu, pengalaman, dan membentuk karakter selama masa studi. Semoga karya ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang machine learning, sistem informasi kesehatan, serta prediksi risiko diabetes berbasis data SIMRS.

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala limpahan rahmat, karunia, dan hidayah-Nya sehingga penulis dapat menyelesaikan penelitian dan penyusunan tesis yang berjudul "Pengembangan Model Prediksi Diabetes Berbasis Hybrid Ensemble dengan Integrasi SMOTE untuk Optimasi Skrining di Sistem Informasi Manajemen Rumah Sakit" ini dengan baik. Penulisan tesis ini merupakan salah satu syarat untuk menyelesaikan studi pada Program Studi Magister Ilmu Komputer/Teknik Informatika.

Penelitian ini dilatarbelakangi oleh keprihatinan terhadap masih tingginya angka underdiagnosis diabetes di Indonesia, di mana sekitar 70% kasus diabetes tidak terdeteksi hingga stadium lanjut. Di sisi lain, data klinis dasar yang tercatat dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS) belum dimanfaatkan secara optimal untuk membantu deteksi dini. Pendekatan hybrid ensemble machine learning yang dikembangkan dalam penelitian ini diharapkan dapat menjadi solusi inovatif yang berkontribusi nyata dalam upaya pencegahan dan pengendalian diabetes di Indonesia.

Proses penyusunan tesis ini tidak terlepas dari bantuan, bimbingan, dukungan, dan doa dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis menyampaikan ucapan terima kasih dan penghargaan yang setinggi-tingginya kepada:

1 Prof. Dr. Ema Utami, S.Si., M.Kom selaku pembibing tesis.
2. Dhani Ariatmanto, S.Kom., M.Kom., Ph.D. selaku penguji 2

3 Alva Hendi Muhammad, A.Md., S.T., M.Eng., Ph.D. selaku penguji 1
Dan tentunya terimakasih kepada ibu saya Ratna kilian

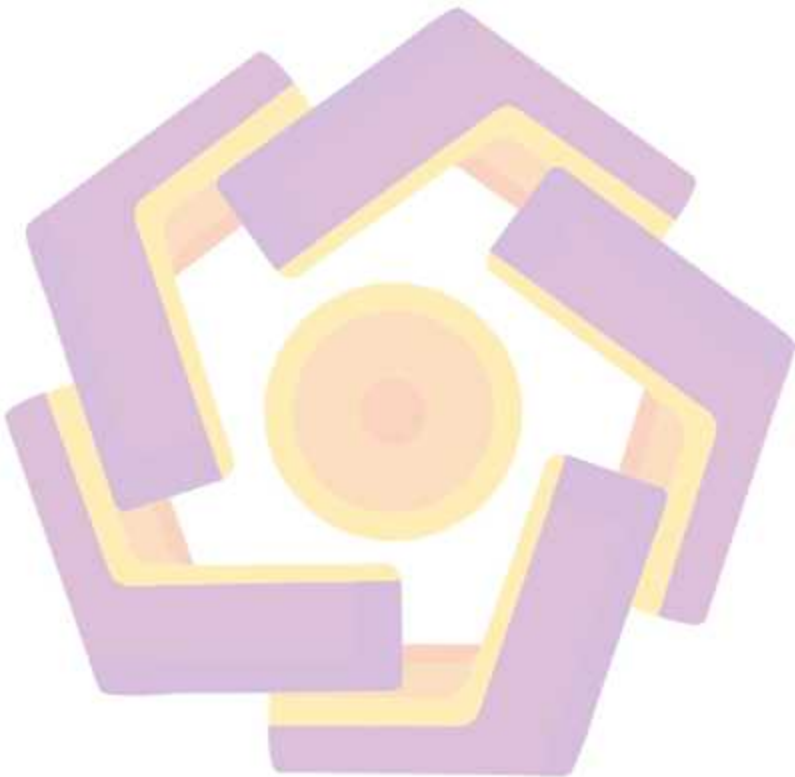
Penulis menyadari bahwa tesis ini masih jauh dari sempurna dan memiliki keterbatasan. Oleh karena itu, penulis dengan terbuka menerima kritik dan saran yang membangun dari berbagai pihak demi perbaikan dan pengembangan penelitian selanjutnya. Akhir kata, semoga tesis ini dapat memberikan manfaat dan kontribusi nyata bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan machine learning untuk prediksi penyakit kronis di Indonesia, serta bagi peningkatan kualitas pelayanan kesehatan dan kesejahteraan masyarakat.

Akhirul kalam, semoga segala kebaikan dan bantuan yang telah diberikan mendapat balasan yang berlipat ganda dari Tuhan Yang Maha Esa.

Yogyakarta, 11 feb 2026

Muh ikbal sodikin

DAFTAR ISI



DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

Ω	Tahanan Listrik
μ	Konstanta gesekan
ANFIS	Adaptive Network Fuzzy Inference System

DAFTAR ISTILAH

Vektor	besaran yang mempunyai arah
Eigen Value	akar akar persamaan

INTISARI

Diabetes mellitus merupakan masalah kesehatan masyarakat Indonesia yang semakin meningkat dengan prevalensi 10,9% pada populasi dewasa. Tantangan utama dalam sistem layanan kesehatan adalah fenomena underdiagnosis, di mana sekitar 70% kasus diabetes tidak terdiagnosis hingga stadium lanjut. Penelitian ini bertujuan mengembangkan model prediksi diabetes yang akurat dengan memanfaatkan data klinis dasar dari Sistem Informasi Manajemen Rumah Sakit (SIMRS) melalui pendekatan hybrid ensemble machine learning.

Metode penelitian menggunakan data retrospektif dari SIMRS sebanyak 958 sampel dengan 7 variabel klinis awal. Tahapan penelitian meliputi: (1) pra-pemrosesan data dan validasi klinis, (2) advanced feature engineering yang mengembangkan 7 fitur menjadi 23 fitur melalui transformasi polynomial, clinical flags, dan interaction terms, (3) penanganan ketidakseimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE), (4) pengembangan model hybrid ensemble yang mengkombinasikan XGBoost, LightGBM, Random Forest, dan Gradient Boosting dengan mekanisme weighted voting, serta (5) optimasi threshold klasifikasi.

Hasil penelitian menunjukkan bahwa proses feature engineering berhasil meningkatkan kapasitas prediktif dataset dengan peningkatan mutual information sebesar 87,5%. SMOTE efektif menangani ketidakseimbangan kelas ekstrem (rasio 1:4.04) dengan meningkatkan distribusi kelas minoritas dari 19,8% menjadi 33,3%. Model hybrid ensemble dengan konfigurasi bobot optimal (XGBoost: 0.40, LightGBM: 0.30, Random Forest: 0.20, Gradient Boosting: 0.10) mencapai akurasi **82,64%**, mengungguli semua model individual. Performa model meningkat signifikan dengan recall **51,72%** (+6,89% dari XGBoost individual), precision **57,69%** (+9,54%), dan F1-score **54,55%** (+8,12%). Optimasi threshold menemukan nilai optimal pada **0,504** (bukan 0,5 konvensional) yang memberikan balance terbaik antara precision dan recall untuk aplikasi skrining klinis.

Penelitian ini memberikan kontribusi dalam pengembangan arsitektur weighted voting hybrid ensemble yang mengoptimalkan kombinasi empat algoritma machine

learning, integrasi SMOTE dengan hybrid ensemble untuk data tidak seimbang, dan metodologi threshold optimization untuk aplikasi klinis. Model yang dikembangkan berpotensi diimplementasikan sebagai sistem pendukung keputusan klinis untuk skrining diabetes di fasilitas pelayanan kesehatan Indonesia, dengan rekomendasi untuk validasi multi-center dan integrasi dengan workflow klinis yang ada.

Kata kunci: Prediksi Diabetes, Hybrid Ensemble, SMOTE, XGBoost, Sistem Informasi Manajemen Rumah Sakit, Machine Learning, Weighted Voting, Feature Engineering, Optimasi Threshold, Skrining Klinis.

ABSTRACT

Diabetes mellitus is an increasing public health problem in Indonesia with a prevalence of 10.9% in the adult population. The main challenge in the healthcare system is the phenomenon of underdiagnosis, where approximately 70% of diabetes cases remain undiagnosed until advanced stages. This study aims to develop an accurate diabetes prediction model by utilizing basic clinical data from Hospital Management Information Systems (SIMRS) through a hybrid ensemble machine learning approach.

The research method uses retrospective data from SIMRS consisting of 958 samples with 7 initial clinical variables. The research stages include: (1) data preprocessing and clinical validation, (2) advanced feature engineering that develops 7 features into 23 features through polynomial transformations, clinical flags, and interaction terms, (3) handling class imbalance using Synthetic Minority Over-sampling Technique (SMOTE), (4) development of a hybrid ensemble model combining XGBoost, LightGBM, Random Forest, and Gradient Boosting with a weighted voting mechanism, and (5) classification threshold optimization.

The results show that the feature engineering process successfully increased the predictive capacity of the dataset with an 87.5% improvement in mutual information. SMOTE effectively handled extreme class imbalance (ratio 1:4.04) by increasing the minority class distribution from 19.8% to 33.3%. The hybrid ensemble model with optimal weight configuration (XGBoost: 0.40, LightGBM: 0.30, Random Forest: 0.20, Gradient Boosting: 0.10) achieved an accuracy of 82.64%, outperforming all individual models. Model performance improved significantly with recall of 51.72% (+6.89% from individual XGBoost), precision of 57.69% (+9.54%), and F1-score of 54.55% (+8.12%). Threshold optimization found

the optimal value at 0.504 (not the conventional 0.5) which provides the best balance between precision and recall for clinical screening applications.

This research contributes to the development of a weighted voting hybrid ensemble architecture that optimizes the combination of four machine learning algorithms, integration of SMOTE with hybrid ensemble for imbalanced data, and threshold optimization methodology for clinical applications. The developed model has the potential to be implemented as a clinical decision support system for diabetes screening in Indonesian healthcare facilities, with recommendations for multi-center validation and integration with existing clinical workflows.

Keywords: Diabetes Prediction, Hybrid Ensemble, SMOTE, XGBoost, Hospital Management Information System, Machine Learning, Weighted Voting, Feature Engineering, Threshold Optimization, Clinical Screening.

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus merupakan penyakit metabolik kronis yang menjadi beban kesehatan masyarakat global, termasuk Indonesia. Berdasarkan data Riset Kesehatan Dasar (Riskesdas) tahun 2018, prevalensi diabetes di Indonesia mencapai 10,9% pada populasi dewasa, meningkat signifikan dari 6,9% pada tahun 2013 [1]. Fenomena underdiagnosis yang serius terjadi, dengan estimasi bahwa sekitar 70% kasus diabetes tidak terdiagnosis hingga stadium lanjut, berpotensi menyebabkan komplikasi mikrovaskular dan makrovaskular yang menurunkan kualitas hidup serta meningkatkan beban biaya kesehatan [2].

Dalam sistem layanan kesehatan primer Indonesia, implementasi skrining diabetes yang efektif menghadapi kendala struktural yang kompleks. Temuan Kementerian Kesehatan Republik Indonesia (2022) mengungkapkan bahwa kapasitas fasilitas pelayanan kesehatan primer dalam melakukan skrining diabetes masih terbatas, dengan hanya sekitar 35% yang memiliki mekanisme skrining terstruktur, sementara 60% kasus diabetes baru terdiagnosis setelah manifestasi komplikasi klinis [3]. Pemeriksaan laboratorium konvensional seperti pengukuran HbA1c dan oral glucose tolerance test sering kali tidak tersedia secara luas atau tidak terjangkau secara finansial untuk program skrining populasi skala besar.

Secara paradoksal, data klinis dasar pasien—meliputi parameter vital, karakteristik demografis, dan riwayat farmakoterapi—terakumulasi secara rutin dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS) namun belum terintegrasi secara optimal dalam algoritma penilaian risiko diabetes. Data klinis longitudinal ini memiliki potensi prediktif yang signifikan jika dikurasi dan dianalisis melalui pendekatan data mining dan machine learning yang canggih.

Kemajuan teknologi machine learning telah membuka paradigma baru dalam prediksi penyakit kronis. Algoritma pembelajaran mesin memiliki kemampuan untuk mengidentifikasi pola non-linear dan interaksi kompleks antar variabel klinis yang tidak dapat diobservasi melalui analisis statistik konvensional. Namun, implementasinya dalam konteks prediksi diabetes dihadapkan pada dua tantangan metodologis utama: pertama, distribusi data yang tidak seimbang (imbalanced dataset) dengan representasi kasus positif yang minimal; kedua, kebutuhan akan model yang presisi meskipun hanya menggunakan parameter klinis dasar yang tersedia dalam sistem informasi kesehatan.

Teknik Synthetic Minority Over-sampling Technique (SMOTE) telah diakui sebagai solusi efektif untuk masalah ketidakseimbangan data dalam domain medis. Metode ini beroperasi dengan prinsip interpolasi linier untuk membangkitkan sampel sintetik kelas minoritas (positive cases) dalam feature space, sehingga meningkatkan kemampuan klasifikasi tanpa menginduksi overfitting yang melekat pada metode oversampling konvensional.

Penelitian sebelumnya oleh Chen et al. (2023) dalam Diabetes Care berhasil mencapai Area Under Curve (AUC) sebesar 0,89 dengan menerapkan deep learning untuk prediksi diabetes [4]. Namun, arsitektur model tersebut memerlukan data longitudinal komprehensif yang mencakup parameter laboratorium, yang sering kali tidak tersedia dalam database SIMRS fasilitas pelayanan kesehatan primer. Di sisi lain, studi Kumar et al. (2023) dalam Journal of Biomedical Informatics mendemonstrasikan bahwa teknik feature engineering yang canggih dapat mengekstrak informasi prediktif maksimal dari variabel klinis dasar [5].

Pendekatan hybrid ensemble merepresentasikan solusi inovatif untuk mengatasi keterbatasan algoritma tunggal. Framework ini mengintegrasikan beberapa algoritma machine learning—termasuk Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Random Forest, dan Gradient Boosting Machine—dalam satu arsitektur terpadu dengan mekanisme weighted voting. Setiap algoritma berkontribusi berdasarkan keunggulan komputasionalnya: XGBoost dengan regularisasi L1/L2 yang kuat, LightGBM

dengan efisiensi pemrosesan data kategorikal, Random Forest dengan robustness terhadap overfitting, dan Gradient Boosting dengan optimasi loss function bertahap.

Sintesis ensemble ini memungkinkan generalisasi model yang lebih baik dan stabilitas prediksi yang lebih tinggi dibandingkan pendekatan unimodal. Fleksibilitas sistem juga memungkinkan adaptasi terhadap heterogenitas karakteristik data antar institusi kesehatan, suatu faktor kritis mengingat variasi dalam implementasi SIMRS di berbagai fasilitas pelayanan kesehatan di Indonesia.

Berdasarkan analisis gap penelitian dan kebutuhan klinis ini, penelitian ini bertujuan mengembangkan model prediksi diabetes berbasis hybrid ensemble dengan integrasi SMOTE yang dioptimalkan untuk data klinis dasar dalam sistem SIMRS. Model ini dirancang untuk mengatasi masalah ketidakseimbangan data sekaligus memaksimalkan akurasi prediksi menggunakan parameter klinis rutin, sehingga berpotensi meningkatkan efisiensi skrining diabetes dan kontribusi pada strategi pencegahan komplikasi diabetes di tingkat nasional.

1.2 Rumusan Masalah

1. Berdasarkan analisis terhadap permasalahan deteksi dini diabetes di fasilitas pelayanan kesehatan Indonesia dan potensi pemanfaatan data rekam medis elektronik, penelitian ini merumuskan tiga pertanyaan penelitian utama:
2. Sejauh mana karakteristik dan kualitas data rekam medis elektronik dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS) dapat dimanfaatkan untuk keperluan prediksi risiko diabetes, serta kendala yang muncul dalam proses *preprocessing* data?
3. Tingkat performa masing-masing algoritma *machine learning* (XGBoost, LightGBM, Random Forest, Gradient Boosting, dan Logistic Regression) dalam memprediksi risiko diabetes berdasarkan data klinis dasar?
4. Efektivitas penerapan model *hybrid ensemble* yang mengombinasikan beberapa algoritma *machine learning* dalam meningkatkan akurasi prediksi risiko diabetes dibandingkan dengan penggunaan algoritma Tunggal?

1.3 Batasan Masalah

1. Penelitian ini menggunakan data rekam medis elektronik yang bersumber dari Sistem Informasi Manajemen Rumah Sakit (SIMRS) pada satu institusi kesehatan, sehingga hasil penelitian belum mewakili kondisi seluruh fasilitas pelayanan kesehatan.
2. Data yang digunakan dalam penelitian ini terbatas pada tujuh variabel klinis dasar, yaitu jenis kelamin, golongan darah, tekanan darah sistolik, tekanan darah diastolik, denyut nadi, usia, dan jumlah obat.
3. Parameter laboratorium, faktor gaya hidup, riwayat keluarga, serta data antropometri tidak digunakan dalam penelitian ini.
4. Algoritma *machine learning* yang digunakan dalam penelitian ini meliputi XGBoost, LightGBM, Random Forest, Gradient Boosting, dan Logistic Regression sebagai model pembanding.
5. Teknik penyeimbangan data yang digunakan hanya terbatas pada Synthetic Minority Over-sampling Technique (SMOTE).
6. Klasifikasi yang dilakukan dalam penelitian ini terbatas pada dua kelas, yaitu pasien diabetes dan pasien non-diabetes

1.4 Tujuan Penelitian

1. Mendapatkan model terbaik dari algoritma Hybrid Ensemble untuk memprediksi risiko diabetes secara akurat dengan mengoptimalkan kombinasi XGBoost, LightGBM, Random Forest, dan Gradient Boosting.
2. Mendapatkan metode penyeimbangan data yang paling efektif untuk mengatasi ketidakseimbangan kelas dalam prediksi diabetes menggunakan teknik SMOTE.
3. Mengetahui tingkat akurasi dan error pada hasil prediksi risiko diabetes dengan mengukur accuracy, precision, recall, F1-score, dan ROC-AUC.
4. Menganalisis pengaruh fitur klinis terhadap hasil prediksi risiko diabetes melalui analisis feature importance.

1.5 Manfaat Penelitian

1. Penelitian ini memberikan manfaat akademik berupa peningkatan pemahaman dan pengalaman dalam penerapan metode *machine learning*, khususnya *hybrid ensemble* dengan mekanisme *weighted*

voting, untuk memprediksi risiko diabetes menggunakan data klinis dasar dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS). Hasil penelitian ini juga dapat menjadi referensi metodologis bagi pengembangan penelitian selanjutnya di bidang informatika kesehatan.

2. Penelitian ini memberikan manfaat praktis dalam mendukung upaya deteksi dini risiko diabetes melalui pemanfaatan data klinis rutin tanpa ketergantungan pada pemeriksaan laboratorium yang kompleks dan mahal. Model prediksi yang dikembangkan berpotensi membantu tenaga kesehatan dan masyarakat dalam mengidentifikasi risiko diabetes secara lebih cepat, objektif, dan konsisten, sehingga dapat berkontribusi pada peningkatan kualitas pelayanan kesehatan serta pencegahan komplikasi diabetes secara lebih dini.
3. Penelitian ini diharapkan dapat menjadi dasar bagi pengembangan sistem pendukung pengambilan keputusan klinis (*clinical decision support system*) berbasis SIMRS yang dapat diintegrasikan pada layanan kesehatan di masa mendatang, serta mendukung efisiensi pelayanan dan pengambilan keputusan medis berbasis data.
4. Secara jangka panjang, hasil penelitian ini berkontribusi pada peningkatan kesadaran dan upaya pencegahan penyakit diabetes di masyarakat, serta membantu menurunkan beban biaya kesehatan melalui pendekatan skrining dan pencegahan berbasis teknologi informasi.

1.6 SISTEMATIKA PENULISAN

Penulisan tesis ini disusun dalam beberapa bab yang saling berkaitan untuk memberikan gambaran yang sistematis dan terstruktur mengenai penelitian yang dilakukan, dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan latar belakang penelitian yang menjelaskan urgensi deteksi dini diabetes serta potensi pemanfaatan data rekam medis elektronik dalam Sistem Informasi Manajemen Rumah Sakit (SIMRS). Selain itu, bab ini memuat rumusan

masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan sebagai kerangka keseluruhan tesis.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI

Bab ini membahas teori dan konsep yang menjadi dasar penelitian. Pembahasan meliputi:

Konsep diabetes mellitus dan karakteristik klinisnya

Sistem Informasi Manajemen Rumah Sakit (SIMRS) dan data rekam medis elektronik

Data mining dan machine learning dalam bidang kesehatan

Algoritma machine learning yang digunakan (Logistic Regression, Random Forest, Gradient Boosting, XGBoost, dan LightGBM)

Teknik penanganan data tidak seimbang menggunakan SMOTE

Konsep hybrid ensemble dan mekanisme weighted voting

Penelitian terdahulu yang relevan sebagai dasar analisis gap penelitian

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan dan metode penelitian yang digunakan, meliputi:

Desain dan alur penelitian

Sumber dan karakteristik data penelitian

Tahapan preprocessing data

Penerapan teknik penyeimbangan data menggunakan SMOTE

Proses pembentukan model machine learning dan hybrid ensemble

Metode evaluasi model menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen dan analisis yang diperoleh dari penerapan model. Pembahasan mencakup:

Analisis karakteristik data sebelum dan sesudah preprocessing

Perbandingan performa algoritma machine learning tunggal

Evaluasi kinerja model hybrid ensemble

Analisis feature importance untuk mengetahui pengaruh variabel klinis

Pembahasan hasil penelitian dibandingkan dengan penelitian terdahulu

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang diperoleh berdasarkan hasil penelitian serta menjawab rumusan masalah. Selain itu, disampaikan saran untuk pengembangan penelitian selanjutnya dan potensi penerapan model dalam sistem pendukung keputusan klinis berbasis SIMRS.

BAB 2
TINJAUAN PUSTAKA

2.1 Tijauan Pustaka

Penelitian mengenai klasifikasi penyakit diabetes dengan pendekatan machine learning, didorong oleh kebutuhan akan sistem pendukung keputusan klinis yang mampu mengolah data medis ,maka Algoritma Extreme Gradient Boosting (XGBoost) kerap menjadi pilihan utama karena reputasinya dalam hal generalisasi, efisiensi komputasi, dan ketangguhan menangani karakteristik data medis yang variati

Berikut adalah tinjauan terhadap beberapa penelitian terdahulu yang relevan. Wang dkk. [1] melakukan studi komparatif di Beijing untuk mengklasifikasikan risiko diabetes tipe 2 dengan membandingkan XGBoost, Support Vector Machine (SVM),). Hasil penelitian menunjukkan superioritas XGBoost yang mencapai akurasi 89,09% dan AUC 0,9182, mengungguli algoritma pembanding. Temuan ini mengindikasikan kemampuan XGBoost dalam menangani kompleksitas data survei kesehatan dan menghasilkan prediksi yang lebih akurat

Penelitian yang juga di lakukan oleh Abdurrahman dkk. [2] menggarisbawahi pentingnya optimasi hyperparameter dalam meningkatkan kinerja algoritma XGBoost, khususnya pada kasus prediksi diabetes. Karakteristik data rekam medis elektronik pada SIMRS umumnya menunjukkan ketidakseimbangan distribusi kelas antara pasien diabetes dan non-diabetes, serta adanya korelasi yang bervariasi antar fitur klinis seperti usia, indeks massa tubuh, kadar glukosa, dan tekanan darah, yang berperan penting dalam proses klasifikasi risiko diabetes. Setelah diterapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) untuk menangani ketidakseimbangan kelas, performa algoritma XGBoost dalam memprediksi risiko diabetes mengalami peningkatan

yang signifikan, terutama pada metrik evaluasi yang sensitif terhadap kelas minoritas. Penerapan SMOTE terbukti memberikan pengaruh positif terhadap peningkatan nilai Recall dan F1-Score, yang menunjukkan kemampuan model dalam mendeteksi kasus diabetes menjadi lebih baik tanpa mengorbankan performa secara keseluruhan. Dalam perbandingan performa antara algoritma XGBoost dan

Support Vector Machine (SVM) pada data SIMRS yang telah diseimbangkan, XGBoost umumnya menunjukkan hasil yang lebih unggul dan stabil, terutama dalam menangkap hubungan nonlinier antar fitur klinis. Selain itu, analisis feature importance pada model XGBoost mengindikasikan bahwa fitur klinis seperti kadar glukosa darah, indeks massa tubuh, usia, dan riwayat kehamilan merupakan faktor yang paling berpengaruh secara signifikan terhadap hasil prediksi risiko diabetes. Temuan ini sejalan dengan hasil penelitian Abdurrahman dkk. yang menggunakan dataset PIMA, di mana akurasi model XGBoost tanpa tuning hanya mencapai 75%, namun meningkat secara drastis hingga 95% setelah penerapan optimasi hyperparameter menggunakan Grid Search dan Random Search, sehingga menegaskan bahwa kinerja optimal XGBoost sangat bergantung pada strategi optimasi parameter yang diterapkan. Inovasi dalam pendekatan optimasi model

Penelitian berikutnya oleh Bandil dan Dandotiya, dengan judul Hyperparameter Optimization Techniques for Enhanced Diabetes Prediction Using XGBoos [3] yang memperkuat pandangan bahwa kombinasi teknik penyeimbangan data dan optimasi hyperparameter merupakan suatu Solusi dalam membangun model prediksi diabetes yang baik dan akurat.

Penelitian berikut yang di tulis oleh W. Li, Y. Peng dkk dengan judul Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm, [4] dengan pendekatan Teknik penyeimbangan data seperti *Random Oversampling*, *ADASYN*, *SMOTE* untuk mengatasi dominasi kelas non-diabetes. emberikan kinerja terbaik dalam meningkatkan kualitas data latih dan performa model. Untuk meningkatkan akurasi prediksi, penulis mengoptimalkan hiperparameter XGBoost menggunakan Genetic Algorithm (GA), yang terbukti lebih efektif dibandingkan metode *grid search*, *random search*, dan *Bayesian*

optimization. Selanjutnya, model GA-XGBoost dikombinasikan dengan LightGBM dan Random Forest

Gregorius Airlangga dengan judul penelitian Enhancing Diabetes Prediction Accuracy through Hybrid Machine Learning Models: A Comparative Study [5] melakukan studi komparatif terhadap beberapa algoritma *machine learning*, yaitu Decision Tree, Random Forest, K-Nearest Neighbors (KNN), dan XGBoost, serta mengombinasikannya dalam model hibrida menggunakan teknik soft voting dan stacking ensemble. Penelitian tersebut menggunakan dataset klinis yang di ambil dari Kaggle

Hasil penelitian menunjukkan bahwa model hibrida secara konsisten mengungguli model pembelajaran tunggal,

Penelitian Zhou et al. (2023) memberikan kontribusi signifikan dalam pengembangan model prediksi diabetes melalui integrasi seleksi fitur Boruta, klastering K-Means++, dan ensemble stacking. [5] Model ini menunjukkan performa yang kompetitif dan stabil dibanding studi sebelumnya. Namun, masih terdapat peluang pengembangan lebih lanjut, terutama dalam penanganan data tidak seimbang, penggunaan dataset yang lebih besar dan beragam, serta uji coba pada lingkungan klinis nyata.

Berdasarkan tinjauan literatur yang telah dilakukan, dapat ditarik beberapa kesimpulan kritis yang menjadi fondasi bagi penelitian ini ,XGBoost telah terbukti sebagai algoritma yang handal untuk prediksi diabetes dengan kemampuan generalisasi yang baik dan efisiensi komputasi yang tinggi, sehingga cocok untuk diimplementasikan dalam skenario data medis yang kompleks namun membutuhkan kecepatan pengolahan.

Dengan ini , optimasi hiperparameter dan penanganan data tidak seimbang merupakan faktor kritis yang signifikan mempengaruhi performa model. Tanpa strategi penyeimbangan yang sesuai, potensi prediktif dari algoritma sekalipun tidak dapat tercapai secara optimal, maka di temukan pendekatan ensemble dan hybrid semakin banyak diterapkan untuk meningkatkan akurasi dan stabilitas

prediksi, menunjukkan tren bahwa model tunggal seringkali belum cukup untuk menangkap kompleksitas pola dalam data Kesehatan.

integrasi teknik pra-pemrosesan yang komprehensif—meliputi seleksi fitur, klastering, dan penyeimbangan data—masih menjadi area yang dapat dikembangkan lebih lanjut, mengingat tahap ini sangat menentukan kualitas data yang menjadi masukan bagi model.

Dengan mempertimbangkan temuan-temuan tersebut, penelitian ini bertujuan untuk mengembangkan model prediksi diabetes yang tidak hanya mengandalkan kekuatan XGBoost sebagai fondasi, tetapi juga mengintegrasikan tahap pra-pemrosesan data yang sistematis, penanganan ketidakseimbangan kelas dengan metode yang lebih adaptif, serta optimasi hiperparameter yang lebih canggih melalui pendekatan automated machine learning atau algoritma. Dengan demikian, diharapkan model yang dihasilkan tidak hanya mencapai kinerja metrik yang tinggi, tetapi juga lebih robust, interpretable, dan siap diterapkan dalam konteks klinis yang sesungguhnya, menjawab kebutuhan akan sistem pendukung keputusan yang andal dan efisien di lingkungan pelayanan kesehatan.

2.2 Keaslian Penelitian

Tabel 2.1 Matriks literatur review dan posisi penelitian

PENGEMBANGAN MODEL PREDIKSI DIABETES BERBASIS HYBRID ENSEMBLE DENGAN INTEGRASI SMOTE UNTUK OPTIMASI SKRINING DI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT

No	JudUL Penelitian	Nama Peneliti, Tahun, Index	Metode Penelitian	Hasil	Keunggulan dan Kelemahan	Perbandingan
1	Prediction of Type 2 Diabetes Risk and Its Effect Using Machine Learning Algorithms	Wang et al., *IJERPH*, 2020 https://doi.org/10.3390/ijerph17062012	Membandingkan XGBoost, SVM, RF, KNN dalam klasifikasi diabetes tipe 2.	XGBoost unggul (akurasi 89,09%; AUC 0,9182).	Sampel terbatas (survei), bukan data rumah sakit.	Penelitian ini menggunakan data SIMRS yang lebih kompleks, bukan survei.
2	Optimasi Algoritma XGBoost Classifier Menggunakan GridSearch dan Random Search	Abdurrahman et al., *JTIIK*, 2022 https://jtiik.ub.ac.id/index.php/jtiik/a	Mengoptimasi hyperparameter XGBoost.	Akurasi meningkat dari 75% → 95% setelah tuning.	Dataset kecil (PIMA), tidak mencerminkan data klinis riil.	Penelitian ini memakai data SIMRS dan fokus pada klasifikasi pasien nyata.

	untuk Prediksi Diabetes	rticle/view/6879				
3	Hyperparameter Optimization Techniques for Enhanced Diabetes Prediction Using XGBoos	Bandil & Dandotiya, *IJRITCC*, 2023 http://www.ijritcc.org/abstract.php?id=9340	Mengintegrasikan Whale Optimization Algorithm (WOA) untuk tuning XGBoost.	WOA-XGBoost meningkatkan akurasi dibanding default.	Evaluasi terbatas pada PIMA dataset	Penelitian ini menguji optimasi pada data SIMRS, bukan dataset publik.
4	Diabetes Prediction Model Based on GA-XGBoost and Stacking Ensemble Learning	Li et al., *PLOS ONE*, 2024 https://journals.plos.org/plosone/article?id=10.1371/journal.	Mengembangkan GA-XGBoost dengan stacking ensemble untuk prediksi diabetes.	GA-XGBoost meningkatkan akurasi	Komputasi kompleks, butuh resource tinggi.	Penelitian ini lebih fokus pada penerapan dengan kompleksitas data lokal.

		pone.03047 16				
5	Enhancing Diabetes Prediction Accuracy through Hybrid Machine Learning Models: A Comparative Study	Gregorius Airlangga, *G-Tech : Jurnal Teknologi Terapan Vol. 8, No. 2, April 2024, pp. 1297-1306 E-ISSN: 2623-064X | P-ISSN: 2580-8737*	Hybrid Models (gabungan RF + Boosting + Feature Selection)	Hybrid model meningkatkan akurasi dibanding single model tradisiona	Performa lebih tinggi karena kombinasi algoritma kompleksitas tinggi dan susah di gunakan	Hybrid > Random Forest tunggal > Logistic Regression
6	Optimalisasi Model Klasifikasi Diabetes Menggunakan	Wibisono et al., *JTSiskom,*	Membandingk an AdaBoost, Gradient	Gradient Boosting unggul	Fitur dataset berbeda dari rekam medis rumah sakit	Penelitian ini lebih relevan karena langsung

	AdaBoost, Gradient Boosting, dan XGBoost	2024 https://jtsiskom.ub.ac.id/index.php/jtsiskom/article/view/543	Boosting, dan XGBoost.	akurasi, XGBoost efisien.		menggunakan data pasien SIMRS.

Source: Author (2007). Gunakan style Citation for Table AMIKOM

2.3 Landasan Teori

Penelitian ini berfokus pada pengembangan model prediksi diabetes menggunakan data dari Sistem Informasi Manajemen Rumah Sakit (SIMRS). SIMRS adalah sistem terintegrasi yang mencatat seluruh aktivitas klinis dan administratif rumah sakit, termasuk data demografis, klinis, laboratorium, dan terapi pasien. Dalam konteks prediksi diabetes, SIMRS menyediakan data

Data SIMRS memiliki karakteristik unik yang berbeda dari dataset standar yang biasa digunakan dalam kompetisi machine learning:

1. Heterogenitas Format: Data dimasukkan oleh berbagai tenaga medis dengan latar belakang berbeda, mengakibatkan variasi dalam format dan kelengkapan data.

2. Ketidaklengkapan Data (Missing Values): Tidak semua variabel diisi secara konsisten untuk setiap pasien, terutama variabel yang tidak langsung terkait dengan diagnosis utama.

3. Ketidakseimbangan Kelas (Class Imbalance): Kasus diabetes (kelas positif) secara signifikan lebih sedikit daripada kasus non-diabetes (kelas negatif), mengikuti prevalensi penyakit dalam populasi umum.

4. Variabilitas Pengukuran Klinis: Pengukuran seperti tekanan darah dan nadi dapat bervariasi berdasarkan kondisi pengukuran dan alat yang digunakan.

Prediksi diabetes dalam penelitian ini didefinisikan sebagai masalah klasifikasi biner, dimana model harus memetakan vektor fitur klinis $x \in r^d$ ke dalam kelas $y \in \{0,1\}$ dengan 0 menunjukkan non-diabetes dan 1 menunjukkan diabetes. Model ini bertujuan untuk memperkirakan probabilitas bersyarat P (y=1|x) yang kemudian dapat digunakan untuk identifikasi dini pasien berisiko.

2.3.1 Alur Kerja Hybrid Machine Learning untuk Prediksi Diabetes

Penelitian ini mengimplementasikan alur kerja yang terstruktur dan sistematis, terdiri dari delapan tahap utama:

Tahap 1: Data Extraction and Feature Engineering

- Ekstraksi data mentah dari SIMRS

- Konstruksi fitur baru berdasarkan pengetahuan klinis

- Transformasi variabel untuk meningkatkan kemampuan prediktif model

Tahap 2: Data Preprocessing

- Encoding variabel kategorikal menjadi format numerik

- Penanganan missing values dengan metode yang tepat

- Pembersihan data dari nilai yang tidak masuk akal secara klinis

Tahap 3: Data Splitting dengan Stratified Sampling

- Pembagian data menjadi set pelatihan (85%) dan pengujian (15%)

- Pemeliharaan proporsi kelas diabetes/non-diabetes di kedua set

Tahap 4: Data Balancing dengan SMOTE

- Penyeimbangan distribusi kelas di data pelatihan

- Generasi sampel sintetis untuk kelas minoritas (diabetes)

Tahap 5: Multi-Model Training

- Pelatihan empat algoritma berbeda secara independen

- Setiap algoritma dikonfigurasi dengan hyperparameter optimal

Tahap 6: Ensemble Construction dan Weight Optimization

- Pembentukan ensemble dengan menggabungkan prediksi individual

- Optimasi bobot setiap model dalam ensemble

Tahap 7: Threshold Optimization

- Penentuan threshold klasifikasi optimal (bukan menggunakan 0.5 default)

- Optimasi berdasarkan trade-off antara sensitivitas dan spesifisitas

Tahap 8: Comprehensive Evaluation dan Visualization

- Evaluasi kinerja dengan multiple metrics

- Visualisasi hasil untuk interpretasi yang lebih baik

Alur ini didesain untuk memaksimalkan akurasi prediksi sambil mempertahankan interpretabilitas klinis.

2.3.2 Feature Engineering Teoritis

Feature engineering adalah proses kritis yang mengubah data mentah menjadi fitur yang lebih informatif untuk algoritma machine learning. Dalam konteks klinis, proses ini harus mempertimbangkan pengetahuan medis yang relevan.

Clinical Flags (Bendera Klinis)

Clinical flags mengkonversi variabel kontinu menjadi variabel biner berdasarkan threshold klinis yang telah divalidasi:

- 'HYPERTENSION_FLAG = (SISTOLE ≥ 140 OR DIASTOLE ≥ 90)': Menandakan pasien dengan hipertensi berdasarkan kriteria JNC 8 (Joint National Committee 8). Threshold 140/90 mmHg merupakan batas diagnosis hipertensi pada orang dewasa.

- 'TACHYCARDIA_FLAG = (NADI > 100)': Menandakan takikardia, dimana denyut nadi >100 denyut per menit dianggap abnormal pada orang dewasa saat istirahat.

- 'POLYPHARMACY_SEVERE = (JUMLAH_OBAT ≥ 8)': Menandakan polifarmasi berat, yang merupakan faktor risiko independen untuk berbagai komplikasi termasuk gangguan kontrol glikemik. Threshold 8 obat dipilih berdasarkan literatur yang menunjukkan peningkatan risiko signifikan pada jumlah ini.

- 'ELDERLY_FLAG = (USIA ≥ 65)': Menandakan usia lanjut, dimana risiko diabetes meningkat secara signifikan. Threshold 65 tahun merupakan standar definisi usia lanjut menurut WHO.

Transformasi Non-linear

Transformasi kuadrat diterapkan pada variabel tertentu untuk menangkap hubungan non-linear:

- 'USIA^', 'SISTOLE^', 'DIASTOLE^', 'JUMLAH_OBAT^': Transformasi ini memungkinkan model untuk menangkap efek kuadratik, misalnya risiko diabetes yang meningkat secara eksponensial dengan usia atau tekanan darah.

Interaction Terms (Istilah Interaksi)

Interaksi antar variabel menangkap efek sinergis:

- 'AGE_BP_RISK = USIA × HYPERTENSION_FLAG': Mengkuantifikasi risiko tambahan dari hipertensi pada kelompok usia tertentu. Secara matematis, ini memungkinkan model untuk memiliki koefisien berbeda untuk hipertensi pada kelompok usia berbeda.

- 'AGE_MED_RISK = USIA × POLYPHARMACY_SEVERE': Menangkap interaksi antara penuaan dan beban pengobatan, yang penting dalam farmakoepidemiologi.

2.3.2 Teori Preprocessing untuk Data Medis

Stratified Sampling Theory

Stratified sampling adalah teknik pembagian data yang mempertahankan proporsi kelas target di setiap subset. Dalam konteks ini:

stratify=y memastikan bahwa proporsi pasien diabetes di set pelatihan sama dengan di set pengujian.

- Rumus: Jika \(p\) adalah proporsi diabetes dalam dataset lengkap, maka stratified sampling menjamin:

$$\frac{\sum y_{train}}{n_{train}} \approx \frac{\sum y_{test}}{n_{test}} \approx p$$

- Manfaat: (1) Estimasi kinerja model yang tidak bias, (2) representasi semua kelas yang memadai di set pengujian, (3) validasi yang lebih reliabel.

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE mengatasi ketidakseimbangan kelas dengan membuat sampel sintetis untuk kelas minoritas:

- Parameter `sampling_strategy=0.5`: Menentukan rasio kelas minoritas terhadap mayoritas setelah resampling. Nilai 0.5 berarti jumlah sampel kelas minoritas akan menjadi 50% dari jumlah sampel kelas mayoritas.

- Algoritma:

1. Untuk setiap sampel x_i di kelas minoritas, temukan k tetangga terdekat (biasanya k=5).

2. Pilih salah satu tetangga secara acak, sebut x_{zi}

3. Hitung perbedaan vektor: $d = x_{zi} - x_i$

4. Pilih bilangan acak λ antara 0 dan 1.

5. Buat sampel sintetis: Xnew=xi+λ·dxnew=$x_i+\lambda \cdot d$.

- Keunggulan dibanding oversampling sederhana: (1) Mencegah overfitting dengan membuat variasi baru, (2) memperluas wilayah keputusan untuk kelas minoritas, (3) menghasilkan sampel yang lebih beragam.

2.3.3 Hybrid Ensemble Theory

A .Weighted Ensemble Theory

Weighted ensemble menggabungkan prediksi dari beberapa model dengan bobot yang berbeda:

- Rumus matematis:

$$P_{ensemble}(y = 1 \mid x) = \sum_{i=1}^{4} w_i \cdot P_{model_i}(y = 1 \mid x), \quad \text{dengan} \sum_{i=1}^{4} w_i = 1 \text{ dan } w_i \geq 0$$

$P_{ensemble}(y = 1 \mid x)$

Probabilitas akhir setelah semua model digabungkan.

$P_{model_i}(y = 1 \mid x)$

Probabilitas yang diprediksi oleh model ke-i

w_i

Bobot model ke-i. Model yang performanya lebih baik bisa diberi bobot lebih besar.

$\sum w_i = 1$

Total bobot harus sama dengan 1 agar hasil tetap dalam rentang probabilitas (0–1).

$w_i \geq 0$

Bobot tidak boleh negatif agar tidak mengurangi makna probabilitas

Daripada mengandalkan satu model saja, pendekatan ini Menggabungkan beberapa model. Memberikan kontribusi berbeda pada masing-masing model.

- Proses optimasi: Bobot dioptimalkan melalui grid search empiris pada empat konfigurasi yang telah ditentukan:

1.[0.4, 0.3, 0.2, 0.1]: Fokus pada XGBoost

2. [0.3, 0.3, 0.2, 0.2]: Distribusi seimbang

3. [0.5, 0.2, 0.2, 0.1]: Dominasi XGBoost yang kuat

4. [0.35, 0.25, 0.25, 0.15]: Distribusi lebih halus

- Kriteria seleksi: Kombinasi bobot yang menghasilkan akurasi tertinggi pada set validasi dipilih.

B .Voting Classifier Theory

Voting classifier adalah teknik ensemble yang menggabungkan prediksi melalui mekanisme voting:

- Soft Voting: Menggabungkan probabilitas prediksi daripada label kelas. Untuk klasifikasi biner:

$$\hat{y} = I\left(\sum_{i=1}^{4} w_i P_i(y = 1 \mid x) \geq 0.5\right)$$

$\hat{y}$ Adalah Hasil prediks kelas 1 atau 0

$$\sum_{i=1}^{4} w_i P_i(y = 1 \mid x) \geq 0.5$$

Probabilitas gabungan (ensemble) yang diperoleh dari penjumlahan probabilitas masing-masing model yang telah dikalikan bobotnya.

w_i ini adalag model bobot

$P_i(y = 1 \mid x)$

Probabilitas model ke-i bahwa data x termasuk kelas positif.

≥ 0.5 Threshold ini Adalah (batas keputusan). Jika probabilitas ≥ 0.5, maka diklasifikasikan sebagai kelas 1.

I(·) Fungsi indikator. Bernilai: 1 jika kondisi di dalam tanda kurung benar dan 0 jika kondisi salah dengan fungsi indikator.

Bobot Voting: Menggunakan bobot optimal yang ditemukan dari weighted ensemble, memastikan konsistensi antara kedua pendekatan ensemble.

C .Threshold Optimization Theory

Threshold klasifikasi menentukan ambang batas probabilitas untuk mengubah prediksi kontinu menjadi label kelas diskrit:

- Rumus optimasi:

$$\tau^* = \arg\max_{\tau \in [0.1,0.9]} \text{Accuracy}\left(y_{test}, I(P(y = 1 \mid x) \geq \tau)\right)$$

τ^* nilai threshod terbaik

$argmax\ \tau \in [0.1,0.9]$ Mencari nilai τ\tauτ dalam rentang 0.1 sampai 0.9 yang memaksimalkan suatu fungsi.

Accuracy Fungsi untuk evaluasi akurasi model.

$ytest$ Label sebenarnya pada data testing.

$I(P(y = 1 \mid x) \geq \tau)$ Hasil klasifikasi berdasarkan threshold τ\tauτ.

Mencoba 100 threshold berbeda secara merata dalam interval [0.1, 0.9], memilih yang menghasilkan akurasi tertinggi.

- Signifikansi klinis: Threshold optimal tidak selalu 0.5. Dalam konteks klinis, threshold yang lebih rendah dapat meningkatkan sensitivitas (mendeteksi lebih banyak kasus benar) dengan mengorbankan spesifisitas (lebih banyak false positive).

2.3.4 Configuration Teoritis Algoritma

A .XGBoost Parameter Configuration

XGBoost (Extreme Gradient Boosting) dikonfigurasi dengan parameter yang dioptimalkan:

- `n_estimators=300`: Jumlah pohon dalam ensemble boosting. Nilai ini dipilih sebagai trade-off antara kinerja dan waktu komputasi.

- `max_depth=6`: Kedalaman maksimal setiap pohon. Membatasi kedalaman mencegah overfitting dan meningkatkan generalisasi.

- `learning_rate=0.05`: Learning rate kecil memberikan konvergensi yang lebih stabil namun memerlukan lebih banyak iterasi.

- `scale_pos_weight`: Dihitung sebagai $\frac{jumlah\ kelas\ negatif}{jumlah\ kelas\ positif}$ di data yang sudah diseimbangkan. Ini memberikan bobot lebih besar pada kesalahan klasifikasi kelas minoritas.

- `subsample=0.8`, `colsample_bytree=0.8`: Mengurangi overfitting dengan menggunakan subset data dan fitur untuk setiap pohon.

- Regularization:

  - `gamma=0.1`: Minimum loss reduction required untuk membuat split

  - `reg_alpha=0.1`: L1 regularization pada leaf weights

  - `reg_lambda=1.5`: L2 regularization pada leaf weights

Konfigurasi Model Lain

- LightGBM: Menggunakan `n_estimators=200`, `max_depth=5`, dan `num_leaves=31`. LightGBM dirancang untuk efisiensi dengan histogram-based algorithm.

- Random Forest: Menggunakan `n_estimators=200` dan `max_depth=8`. Random Forest mengurangi varians melalui bagging dan feature randomness.

- Gradient Boosting: Menggunakan `n_estimators=150` dan `max_depth=5`. Implementasi scikit-learn dari gradient boosting.

2.3.5 Evaluation Framework Theory

Multi-Metric Assessment

Evaluasi model dilakukan dengan lima metrik utama:

1. Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP (True Positive) → Data positif yang diprediksi positif
- TN (True Negative) → Data negatif yang diprediksi negatif
- FP (False Positive) → Data negatif yang salah diprediksi positif
- FN (False Negative) → Data positif yang salah diprediksi negatif

Mengukur proporsi klasifikasi benar secara keseluruhan.

2. Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- TP (True Positive) → Data positif yang diprediksi positif
- FP (False Positive) → Data negatif yang salah diprediksi positif

Mengukur ketepatan prediksi positif (seberapa sering prediksi diabetes benar).

3. Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN}$$

- TP (True Positive) → Data positif yang diprediksi positif
- FN (False Negative) → Data positif yang salah diprediksi negatif

Mengukur kemampuan mendeteksi kasus positif (seberapa banyak kasus diabetes terdeteksi).

4. F1-Score:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Jika Precision tinggi tetapi Recall rendah → F1 tetap rendah
- Jika Recall tinggi tetapi Precision rendah → F1 tetap rendah
- F1 tinggi hanya jika keduanya tinggi

Rata-rata harmonik precision dan recall, berguna ketika kelas tidak seimbang.

5. ROC-AUC:

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR$$

- **AUC** → Luas area di bawah kurva ROC
- **TPR (True Positive Rate)** → Recall
- **FPR (False Positive Rate)** → False Positive Rate
- $\int 01 \int_0^1 \int 01$ → Integral dari 0 sampai 1 (menghitung luas area)

Mengukur kemampuan model membedakan antara kelas di semua threshold yang mungkin.

6.Confusion Matrix Clinical Interpretation

Confusion matrix memberikan wawasan mendalam tentang jenis kesalahan model:

- True Positive (TP): Pasien diabetes yang terdeteksi dengan benar. Implikasi klinis: Intervensi tepat waktu dapat diberikan.

- False Negative (FN): Pasien diabetes yang tidak terdeteksi. Implikasi klinis: Risiko komplikasi meningkat karena tidak ada intervensi.

- False Positive (FP): Pasien non-diabetes yang diprediksi diabetes. Implikasi klinis: Overdiagnosis, biaya tambahan untuk tes konfirmasi, kecemasan pasien.

- True Negative (TN): Pasien non-diabetes yang diklasifikasi dengan benar. Implikasi klinis: Efisiensi sumber daya, tidak ada intervensi yang tidak perlu.

2.3.6 Feature Importance Theory

XGBoost menghitung feature importance dengan tiga metrik berbeda:

1. Weight (Frequency):

$$\text{Weight}_f = \frac{S_f}{\sum_{k=1}^{m} S_k}$$

- S_j = jumlah split yang menggunakan fitur ke-j
- m = jumlah total fitur

Mengukur seberapa sering sebuah fitur digunakan dalam pohon keputusan.

2. Gain (Average Improvement):

$$Gain_j = \frac{1}{m}\sum_{t=1}^{m} \Delta L_t(j)$$

Dimana $\Delta L_t(j)$ adalah reduksi loss ketika fitur j/ digunakan untuk split di pohon t. Mengukur kontribusi fitur terhadap peningkatan performa model.

3. Cover:

$$Cover_j = \frac{1}{m}\sum_{t=1}^{m} n_t(j)$$

Dimana $nt(j)$ adalah jumlah sampel yang dipengaruhi oleh split menggunakan fitur j di pohon t. Mengukur seberapa banyak data yang dipengaruhi oleh fitur tersebut.

2.3.7 Model Persistence Theory

Model persistence mengacu pada penyimpanan model terlatih untuk penggunaan di masa depan:

- Struktur Model Package:

```
{
  'xgb_model': model XGBoost terlatih,
  'lgb_model': model LightGBM terlatih,
  'rf_model': model Random Forest terlatih,
  'gb_model': model Gradient Boosting terlatih,
```

```
'voting_model': model Voting Classifier terlatih,
'feature_names': daftar nama fitur,
'best_weights': bobot optimal untuk ensemble,
'best_threshold': threshold klasifikasi optimal,
'final_accuracy': akurasi akhir model.
'metrics_final': semua metrik evaluasi,
'feature_importance': importance setiap fitur,
'training_date': tanggal dan waktu pelatihan
}
```

- Format Penyimpanan: Menggunakan pickle, format serialisasi Python standar.

- Manfaat: (1) Reproduksibilitas, (2) deployment mudah, (3) version control, (4) dokumentasi lengkap model.

2.3.8 Theoretical Contributions of This Approach

Pendekatan dalam penelitian ini memberikan beberapa kontribusi teoretis:

1. Hybridization: Mengkombinasikan multiple gradient boosting algorithms (XGBoost, LightGBM, Gradient Boosting) dengan bagging algorithm (Random Forest) dalam satu framework terpadu.

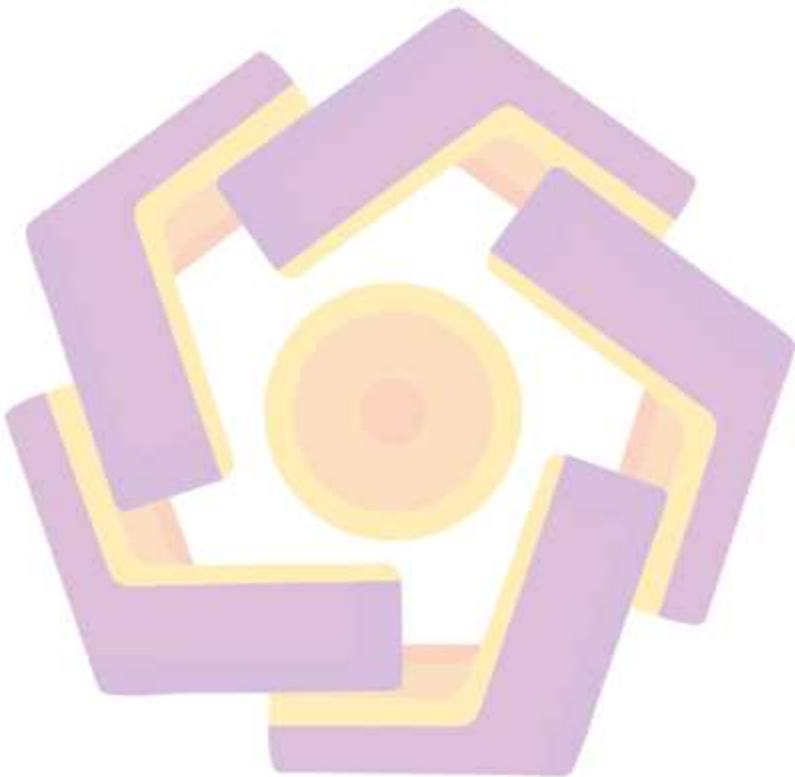
2. Clinical Feature Engineering: Mengintegrasikan pengetahuan klinis mendalam dalam konstruksi fitur, bukan hanya mengandalkan fitur mentah dari dataset.

3. Adaptive Thresholding: Mengoptimalkan threshold klasifikasi berdasarkan data, bukan menggunakan nilai default 0.5, yang meningkatkan performa dalam konteks ketidakseimbangan kelas.

4. Comprehensive Validation: Melakukan evaluasi dengan multiple metrics dan visualisasi komprehensif, memberikan gambaran lengkap tentang kekuatan dan kelemahan model.

5. Deployment Ready: Model dirancang dan disimpan dengan semua komponen yang diperlukan untuk implementasi langsung dalam sistem klinis.

Kontribusi-kontribusi ini memastikan bahwa model tidak hanya memiliki performa statistik yang baik tetapi juga relevan dan dapat diimplementasikan dalam konteks klinis nyata.



BAB 3
METODE PENELITIAN

3.1 Jenis, Sifat, dan Pendekatan Penelitian

3.1.1 Jenis Penelitian

Penelitian ini merupakan penelitian **eksperimental komputasional** dengan pendekatan **kuantitatif** yang bertujuan untuk mengembangkan model prediksi berbasis *machine learning*. Penelitian dikategorikan sebagai **penelitian terapan (*applied research*)** dan **penelitian pengembangan (*development research*)** karena berfokus pada penciptaan solusi praktis—yaitu sistem pendukung keputusan klinis untuk deteksi risiko Diabetes Mellitus tipe 2—dengan memanfaatkan data riil dari lingkungan rumah sakit di Indonesia.

3.1.2 Sifat Penelitian

Penelitian ini memiliki sifat-sifat berikut:

1. **Eksperimental Kuantitatif:** Menggunakan pendekatan kuantitatif dengan desain eksperimen *machine learning* yang terstruktur untuk menguji kinerja berbagai konfigurasi model.
2. **Retrospektif Analitik:** Menganalisis data historis rekam medis elektronik yang dikumpulkan secara retrospektif dari periode waktu tertentu.
3. **Komparatif-Evaluatif:** Membandingkan dan mengevaluasi performa model sebelum dan setelah optimasi serta penanganan masalah ketidakseimbangan data.
4. **Validatif:** Memvalidasi kehandalan dan keakuratan model melalui teknik *cross-validation* dan pengujian pada dataset independen.

3.1.3 Pendekatan Penelitian

Pendekatan penelitian mengintegrasikan dua kerangka utama:

1. **Pendekatan *CRISP-DM* (Cross-Industry Standard Process for Data Mining):** Penelitian mengikuti fase-fase standar CRISP-DM (*Business*

Understanding, Data *Understanding*, ***Data*** *Preparation*, *Modeling*, *Evaluation*, *Deployment*) untuk memastikan metodologi yang sistematis dan terukur.

2. **Pendekatan *Design Science Research*:** Penelitian berorientasi pada perancangan dan penciptaan (model prediksi) sebagai solusi untuk masalah praktis yang diidentifikasi (kesenjangan pemanfaatan data SIMRS), kemudian melakukan evaluasi terhadap artefak tersebut.

3.2 Metode Pengumpulan Data

3.2.1 Sumber Data

Penelitian ini menggunakan data sekunder eksklusif yang berasal dari **Sistem Informasi Manajemen Rumah Sakit (SIMRS)** sebuah rumah sakit di Indonesia. Dataset yang digunakan adalah cleaned_patient_resume_new_ver_06_plan.csv, yang berisi rekam medis terstruktur pasien. Penggunaan dataset lokal ini bertujuan agar model dapat mempelajari pola klinis yang spesifik dan relevan dengan karakteristik populasi pasien di Indonesia.

3.2.2 Karakteristik Dataset

- **Periode Pengambilan Data:** Rekam medis
- **Jumlah Sampel Awal: 958** rekam medis pasien.
- **Distribusi Kelas Target (Diagnosis Diabetes):**
    - Kelas **Diabetes (1):** 190 pasien **(19.8%)**
    - Kelas **Non-Diabetes (0):** 768 pasien **(80.2%)**
- **Tingkat Ketidakseimbangan (*Imbalance Ratio*): 1:4.04** (termasuk ketidakseimbangan yang ekstrem).

3.2.3 Kriteria Inklusi dan Eksklusi Sampel

Untuk memastikan kualitas dan relevansi data, diterapkan kriteria sebagai berikut:

- **Kriteria Inklusi:**
    1. Pasien dewasa (usia ≥ 18 tahun).

2. Memiliki rekam medis lengkap dengan diagnosis pasti (Diabetes atau Non-Diabetes) yang tercatat.
3. Memiliki nilai untuk minimal 5 dari 8 fitur klinis utama yang diteliti.

- **Kriteria Eksklusi:**
  1. Data duplikat untuk pasien yang sama dalam periode yang sama.
  2. Rekam medis dengan lebih dari 50% data fitur yang hilang (*missing values*).
  3. Pasien dengan diagnosis Diabetes Mellitus tipe 1 (ICD-10 E10 - E14).

3.2.4 Teknik Sampling dan Pembagian Data

Teknik **total sampling** diterapkan pada data yang memenuhi kriteria dalam periode yang ditentukan. Dataset kemudian dibagi menjadi subset **latih** dan **uji** dengan proporsi **85:15**.

- **Training Set:** 814 sampel (85%), digunakan untuk membangun dan melatih model.
- **Test Set:** 144 sampel (15%), digunakan untuk evaluasi akhir model yang bersifat independen.
- **Teknik Pembagian:** ***Stratified Random Sampling*** untuk menjaga proporsi distribusi kelas Diabetes dan Non-Diabetes yang sama antara data latih dan data uji.

3.2.5 Variabel Penelitian

1 Variabel Dependen (Target)

- **Nama:** DIAGNOSIS_DIABETES
- **Tipe:** Variabel Biner (Kategorikal Nominal)
- **Nilai: 0** = Non-Diabetes, **1** = Diabetes

- **Sumber Label:** Dibentuk secara *rule-based* berdasarkan kode diagnosis **ICD-10 E 10 E11** dalam rekam medis elektronik. Yang di tampilkan di table berikut

Tabel 3.1 8 fitur awal yang digunakan dalam penelitian

DIAGNOSA_ID	SEX	GOL_DARAH	SISTOLE	DIASTOLE	NADI	USIA	JUMLAH_OBAT
E11.4	F	AB	152.0	91.0	83.0	46	13
E11.6	M	B	165.0	92.0	96.0	58	17
E11.6	M	B	104.0	72.0	69.0	65	11
E11.6	M	B	111.0	59.0	82.0	63	11
E14.9	F	B	119.0	82.0	109.0	33	7
...	...	...	...	...	...	...	...
E11.4	M	B	153.0	93.0	70.0	79	9
E11.4	M	O	144.0	79.0	62.0	68	11
E11.4	M	O	120.0	74.0	82.0	63	9
E11.6	M	B	164.0	92.0	90.0	55	9
E11.4	M	A	138.0	69.0	91.0	47	9

Penjelasan Kaltan 7 Fitur Awal dengan Solusi Ensemble Hybrid:

- 7 fitur awal ini **belum cukup optimal** jika digunakan langsung untuk prediksi kompleks seperti diabetes.
- **Strategi Hybrid:** Sebelum memasuki tahap ensemble, dilakukan **advanced feature engineering**.
- **Contoh Fitur Turunan:** Dari SISTOLE dan DIASTOLE dibuat MAP (Mean

Arterial Pressure), PP (Pulse Pressure). Dari USIA dibuat kategori usia (USIA_CAT).

- **Hasil:** 7 fitur awal ini ditransformasi menjadi **23 fitur yang lebih informatif. Setiap algoritma dalam ensemble hybrid (XGBoost, dll) akan dilatih menggunakan set 23 fitur ini**, bukan 7 fitur awal.

2. **Mengapa Ensemble Hybrid Membutuhkan Feature Engineering dari Fitur Awal**

    - **Kekuatan Berbeda:** Setiap algoritma dalam ensemble punya kekuatan berbeda. XGBoost hebat menangani fitur non-linear, sementara model linear butuh fitur yang sudah dinormalisasi.
    - **Meta-Learner yang Lebih Cerdas:** Pada akhirnya, ensemble hybrid (misal dengan **Stacking**) akan memiliki *meta-learner* (misal Logistic Regression) yang belajar mengombinasikan prediksi dari semua *base learner*. Kualitas prediksi *base learner* sangat bergantung pada kualitas input fitur..

3.3 Metode Analsis Data

3.3.1 Instrumen dan Perangkat Lunak

Analisis data dan pemodelan dilakukan menggunakan perangkat lunak berikut:

- **Bahasa Pemrograman:** Python 3.9+
- **Libraries Utama:**
    - pandas, numpy untuk manipulasi dan analisis data.
    - scikit-learn untuk pra-pemrosesan, pembagian data, dan metrik evaluasi.
    - xgboost untuk implementasi algoritma XGBoost.
    - imbalanced-learn untuk teknik penyeimbangan data SMOTE.
    - matplotlib, seaborn untuk visualisasi data dan hasil.
- **Lingkungan Pengembangan:** Jupyter Notebook.

- **Data Source:** File CSV cleaned_patient_resume_new_ver_06_plan.csv.

3.3.2 Tahapan Analisis Data dan Pemodelan

a. **Pra-Pemrosesan Data (*Data Preprocessing*):**

- Pengecekan dan penanganan *missing values*.
- Deteksi dan penanganan *outlier* dengan metode IQR (*Interquartile Range*).
- **Encoding Variabel Kategorikal:** Mengubah variabel SEX dan GOL_DARAH menjadi bentuk numerik menggunakan *Label Encoding*.
- **Normalisasi:** Melakukan normalisasi pada variabel numerik (seperti USIA, SISTOLE) ke skala yang seragam menggunakan *StandardScaler*.

b. **Rekayasa Fitur dan Penyeimbangan Data:**

- **Rekayasa Fitur (*Feature Engineering*):** Membuat fitur turunan/interaksi baru (misal: *Mean Arterial Pressure* dari SISTOLE & DIASTOLE) dari 7 fitur awal, sehingga total fitur menjadi **23**.
- **Penyeimbangan Kelas:** Menerapkan teknik **SMOTE** hanya pada **data latih** untuk mengatasi ketidakseimbangan ekstrem (1:4.04). Setelah SMOTE, data latih menjadi 979 sampel dengan distribusi yang lebih seimbang.

c. **Pemodelan dengan XGBoost:**

- Algoritma **Extreme Gradient Boosting (XGBoost)** digunakan sebagai model dasar.
- Dilakukan optimasi *hyperparameter* (seperti learning_rate, max_depth, n_estimators) menggunakan teknik *Grid Search* dengan 5-Fold Cross Validation pada data latih.

- Model terbaik dari proses tuning kemudian dilatih ulang pada seluruh data latih.

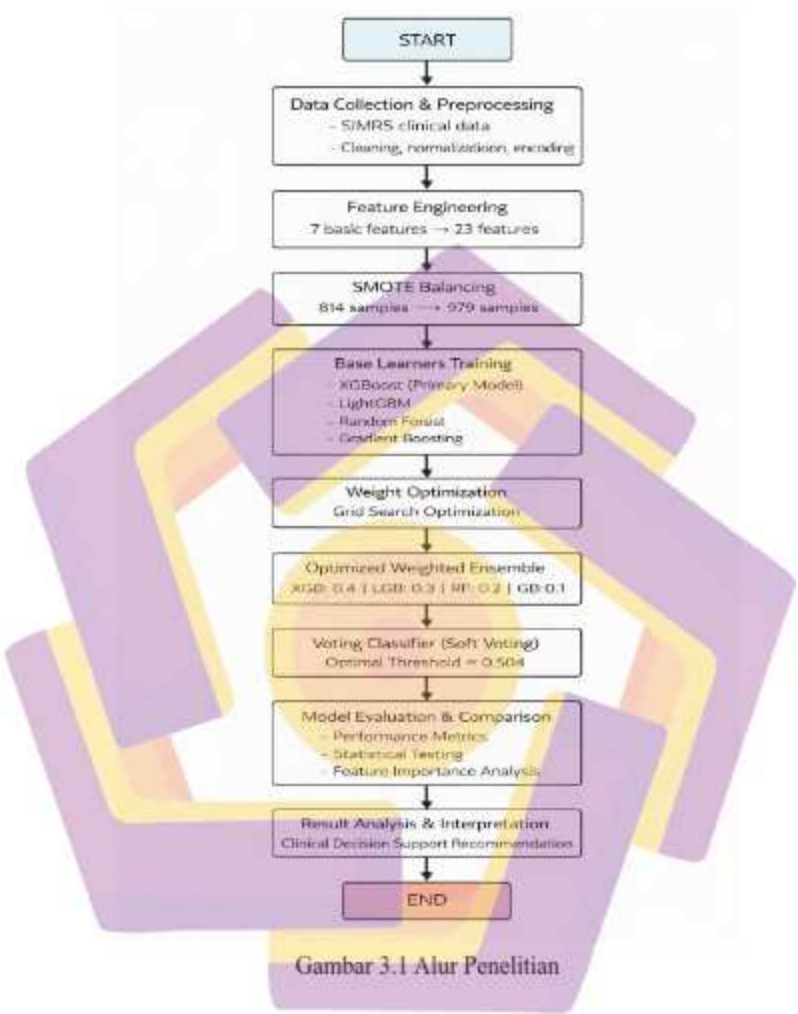
d. **Evaluasi Model:**

- Model dievaluasi secara ketat pada **data uji (144 sampel)** yang tidak tersentuh selama proses tuning dan tidak di-SMOTE.

Berikut alur penelitian memberikan gambaran visual yang jelas mengenai tahapan metodologis pengembangan model ensemble hybrid.

3.3.4 Alur Penelitian Secara Keseluruhan

Diagram ini menggambarkan alur metodologi penelitian dalam pengembangan model prediksi risiko Diabetes Mellitus berbasis machine learning. Proses diawali dengan pengumpulan dan preprocessing data klinis dari SIMRS, dilanjutkan dengan feature engineering untuk memperkaya representasi fitur serta penanganan ketidakseimbangan kelas menggunakan metode SMOTE. Selanjutnya, beberapa model dasar (XGBoost, LightGBM, Random Forest, dan Gradient Boosting) dilatih secara independen dan digabungkan melalui pendekatan weighted ensemble dengan bobot yang dioptimalkan menggunakan grid search. Prediksi akhir dihasilkan menggunakan metode soft voting dengan optimasi nilai threshold untuk memperoleh kinerja klasifikasi yang optimal. Tahap akhir meliputi evaluasi dan perbandingan model berdasarkan metrik kinerja, pengujian statistik, serta analisis pentingnya fitur, yang kemudian diinterpretasikan sebagai dasar penyusunan kesimpulan dan rekomendasi implementasi sistem pendukung keputusan klinis.



Gambar 3.1 Alur Penelitian

a. START

Tahap awal penelitian yang menandai dimulainya proses pengembangan model prediksi risiko Diabetes Mellitus berbasis machine learning.

b. Data Collection & Preprocessing

Pada tahap ini, data klinis pasien dikumpulkan dari **SIMRS** Proses preprocessing dilakukan untuk memastikan kualitas data, meliputi:

- Pembersihan data dari nilai kosong dan duplikasi
- Normalisasi data numerik
- Encoding variable kategorikal tahap ini bertujuan agar data siap digunakan dalam proses pemodelan.

c. Feature Engineering

Fitur-fitur dasar yang diperoleh dari data klinis awal sebanyak **7 fitur** kemudian dikembangkan melalui proses feature engineering. Tahapan ini mencakup:

- Transformasi variabel
- Kombinasi fitur klinis
- Pembuatan fitur turunan yang relevan secara medis Tujuannya adalah meningkatkan representasi informasi dan performa **model.**

d. SMOTE

Balancing Dataset yang tidak seimbang ditangani menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE). Jumlah data meningkat dari 814 sampel menjadi 979 sampel, sehingga distribusi kelas menjadi lebih seimbang. Langkah ini penting untuk mengurangi bias model terhadap kelas mayoritas.

e. Base Learners

Training Pada tahap ini dilakukan pelatihan beberapa model dasar (base learners), yaitu:

- XGBoost sebagai model utama (primary model)
- LightGBM
- Random Forest
- Gradient Boosting

  Setiap model dilatih secara independen untuk mempelajari pola data klinis pasien.

f. Weight Optimization

Bobot masing-masing model dasar dioptimalkan menggunakan **Grid Search Optimization**.

Tujuan tahap ini adalah menentukan kontribusi optimal setiap model dalam ensemble agar kinerja prediksi menjadi maksimal.

g. Optimized Weighted Ensemble

Model-model dasar digabungkan dalam sebuah **weighted ensemble** dengan bobot hasil optimasi, misalnya:

- XGBoost: 0.4
- LightGBM: 0.3
- **Random** Forest: 0.2
- Gradient Boosting: 0.1 Pendekatan ini meningkatkan stabilitas dan akurasi prediksi dibandingkan model tunggal

h. Voting Classifier (Soft Voting)

Prediksi akhir dihasilkan menggunakan soft voting, yaitu dengan menggabungkan probabilitas prediksi dari seluruh model dalam ensemble. Metode ini mempertimbangkan tingkat kepercayaan masing-masing model, bukan hanya hasil klasifikasi akhir.

i. Threshold Optimization

Ambang batas klasifikasi (*decision threshold*) dioptimalkan untuk memperoleh keseimbangan terbaik antara **precision dan recall**.

j. Model Evaluation & Comparison

Kinerja model dievaluasi dan dibandingkan menggunakan:

- **Performance metrics** (accuracy, precision, recall, F1-score, ROC-AUC)
- **Statistical testing** untuk menilai signifikansi perbedaan performa
- **Feature importance analysis** untuk mengidentifikasi fitur klinis yang paling berpengaruh

k. Result Analysis & Interpretation

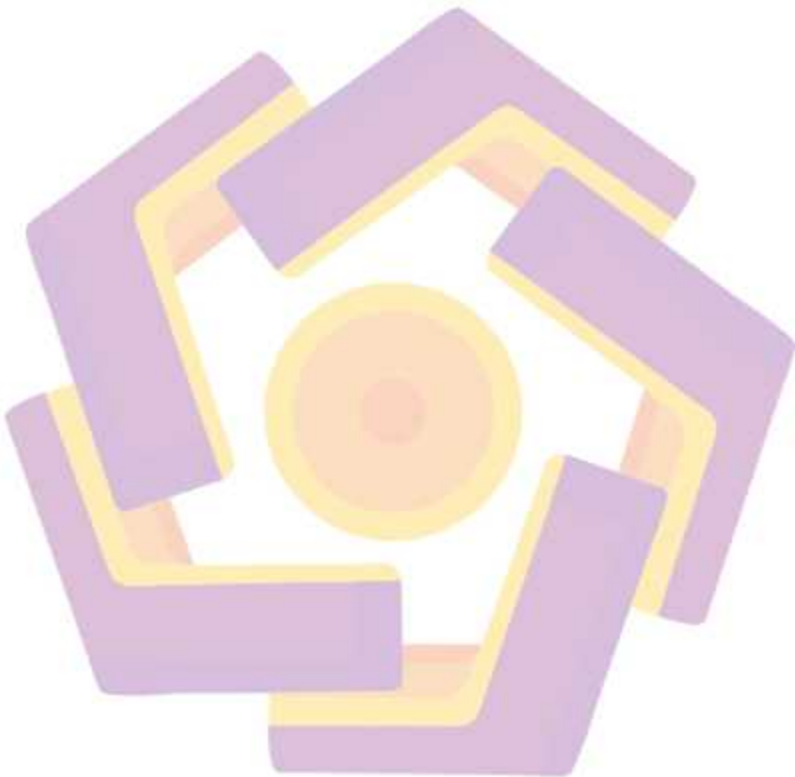
Hasil evaluasi dianalisis dan diinterpretasikan untuk memahami perilaku model serta implikasi klinisnya. Tahap ini bertujuan memastikan bahwa model dapat digunakan sebagai sistem pendukung keputusan klinis yang dapat dipercaya.

l. Conclusion & Implementation

Tahap akhir berisi kesimpulan penelitian serta rekomendasi implementasi model dalam lingkungan klinis, khususnya sebagai alat bantu deteksi dini risiko Diabetes Mellitus.

m. END

Menandai selesainya seluruh rangkaian proses penelitian dan pengembangan model.



BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Deskripsi Komprehensif Dataset dan Validitasnya

Penelitian ini menggunakan basis data 958 rekam medis elektronik pasien yang diperoleh langsung dari Sistem Informasi Manajemen Rumah Sakit (SIMRS) suatu rumah AXZ yang berada magelang Indonesia. Dataset awal terdiri dari 7 variabel klinis utama yang merepresentasikan tiga dimensi data medis: dimensi demografis (usia, jenis kelamin), dimensi klinis-vital (tekanan darah sistolik, diastolik, denyut nadi), dan dimensi farmakologis-administratif (jumlah obat, golongan darah). Dari total sampel, terdapat 190 kasus diabetes (19.8%) dan 768 kasus non-diabetes (80.2%), menunjukkan distribusi kelas yang sangat tidak seimbang

Proses validasi dataset dilakukan melalui pendekatan multi-lapis. Lapis pertama berupa validasi diagnosis melalui pemeriksaan kode ICD-10 yang sesuai dengan standar klasifikasi penyakit internasional, di mana kasus diabetes diidentifikasi berdasarkan kode E

Dataset penelitian ini terdiri atas delapan variabel yang diperoleh dari Sistem Informasi Manajemen Rumah Sakit (SIMRS) dan merepresentasikan karakteristik pasien dari aspek demografis, klinis, farmakologis, serta administratif. Struktur variabel dirancang untuk mencerminkan kondisi pasien secara komprehensif namun tetap sederhana, sehingga relevan untuk pengembangan model prediksi berbasis machine learning. Setiap variabel dipilih berdasarkan pertimbangan klinis serta ketersediaan data rutin pada layanan kesehatan tingkat pertama maupun lanjutan.

Secara konseptual, variabel dalam dataset dikelompokkan ke dalam tiga dimensi utama, yaitu dimensi demografis, dimensi klinis–vital, serta dimensi farmakologis–administratif. Dimensi demografis mencakup usia dan jenis kelamin sebagai faktor risiko dasar yang secara epidemiologis berhubungan dengan kejadian Diabetes Mellitus. Dimensi klinis–vital meliputi tekanan darah sistolik, tekanan darah diastolik, dan denyut nadi, yang menggambarkan kondisi hemodinamik pasien dan

sering berkaitan dengan komorbiditas metabolik. Sementara itu, dimensi farmakologis–administratif mencerminkan kompleksitas terapi melalui jumlah obat yang diresepkan serta karakteristik biologis melalui golongan darah.

Selain tujuh variabel prediktor tersebut, terdapat satu variabel tambahan yaitu DIAGNOSA_ID yang digunakan secara khusus untuk membentuk variabel target penelitian, yakni *is_diabetes*. Variabel ini tidak diikutsertakan sebagai fitur dalam proses pelatihan model guna menghindari kebocoran informasi (*data leakage*). Proses pembentukan target dilakukan melalui fungsi *create_diabetes_target()* dengan mengacu pada kode diagnosis berdasarkan klasifikasi ICD-10, khususnya rentang kode E10–E14 yang merepresentasikan Diabetes Mellitus.

Pada Tabel 4.1 ditampilkan deskripsi rinci setiap variabel prediktor beserta tipe data

Tabel 4.1: Deskripsi dataset Variabel Prediktor

No	Kategori	Nama Variabel	Tipe Data	Deskripsi	Statistik Deskriptif
1	Demografis	**USIA**	Numerik	Usia pasien (tahun)	Mean: 52.3 ± 14.7, Range: 18–89
2	Demografis	**SEX**	Kategorik	Jenis kelamin pasien	Laki-laki: 53.4%, Perempuan: 46.6%
3	Klinis–Vital	**SISTOLE**	Numerik	Tekanan darah sistolik (mmHg)	Mean: 128.5 ± 18.2, Range: 90–190
4	Klinis–Vital	**DIASTOLE**	Numerik	Tekanan darah diastolik (mmHg)	Mean: 82.1 ± 11.5, Range: 60–120
5	Klinis–Vital	**NADI**	Numerik	Denyut nadi istirahat (bpm)	Mean: 76.8 ± 12.4, Range: 50–120
6	Farmakologis	**JUMLAH_OBAT**	Numerik	Jumlah obat yang diresepkan per kunjungan	Mean: 4.2 ± 2.8, Range: 1–15

7	Administratif	**GOL_DARAH**	Kategorik	Golongan darah pasien	A: 30.0%, B: 25.6%, AB: 17.5%, O: 26.9%
8	Klinis–Diagnostik	**DIAGNOSA_ID**	Kategorik	Kode diagnosis utama berdasarkan ICD-10	E10–E14: Diabetes Mellitus, kode lain: Non-Diabetes

ditampilkan deskripsi rinci setiap variabel prediktor beserta tipe data dan statistik deskriptifnya. Rata-rata usia pasien adalah 52,3 ± 14,7 tahun dengan rentang 18 hingga 89 tahun, menunjukkan bahwa dataset didominasi oleh populasi usia dewasa hingga lanjut usia. Proporsi jenis kelamin relatif seimbang dengan komposisi laki-laki sebesar 53,4% dan perempuan 46,6%. Untuk parameter klinis, rerata tekanan darah sistolik tercatat 128,5 ± 18,2 mmHg dan diastolik 82,1 ± 11,5 mmHg, yang mengindikasikan sebagian populasi berada pada kategori prehipertensi hingga hipertensi ringan. Denyut nadi rata-rata sebesar 76,8 ± 12,4 bpm masih berada dalam kisaran fisiologis normal.

Selanjutnya, variabel jumlah obat memiliki rata-rata 4,2 ± 2,8 resep per kunjungan, dengan rentang 1 hingga 15 obat, yang dapat merefleksikan tingkat kompleksitas kondisi klinis pasien. Distribusi golongan darah menunjukkan proporsi terbesar pada golongan A (30,0%), diikuti O (26,9%), B (25,6%), dan AB (17,5%). Secara keseluruhan, karakteristik ini memberikan gambaran awal mengenai struktur data yang digunakan dalam penelitian serta mendukung tahapan analisis lebih lanjut dalam pengembangan model prediksi risiko Diabetes Mellitus.

Dataset penelitian ini terdiri atas delapan variabel yang diperoleh dari Sistem Informasi Manajemen Rumah Sakit (SIMRS) dan merepresentasikan karakteristik pasien dari aspek demografis, klinis, farmakologis, serta administratif. Struktur variabel dirancang untuk mencerminkan kondisi pasien secara komprehensif namun tetap sederhana, sehingga relevan untuk pengembangan model prediksi berbasis machine learning. Setiap variabel dipilih berdasarkan pertimbangan klinis

serta ketersediaan data rutin pada layanan kesehatan tingkat pertama maupun lanjutan.

Tabel 4.2 dengan Model Machine Learning

Jenis Variabel	Digunakan untuk
USIA, SEX, SISTOLE, DIASTOLE, NADI, JUMLAH_OBAT, GOL_DARAH	Fitur Input (X)
DIAGNOSA_ID → DIABETES	Variabel Target (y)

Dalam penelitian ini, tujuh variabel utama yaitu USIA, SEX, SISTOLE, DIASTOLE, NADI, JUMLAH_OBAT, dan GOL_DARAH digunakan sebagai fitur input (X). Variabel-variabel tersebut merepresentasikan kondisi demografis, fisiologis, serta aspek terapi pasien yang secara klinis memiliki potensi keterkaitan dengan risiko Diabetes Mellitus. Seluruh fitur ini diproses melalui tahapan preprocessing seperti encoding untuk data kategorik dan normalisasi atau standarisasi untuk data numerik sebelum dimasukkan ke dalam algoritma pembelajaran mesin.

Sementara itu, variabel DIAGNOSA_ID tidak digunakan secara langsung dalam pelatihan model. Variabel ini terlebih dahulu ditransformasikan menjadi variabel biner bernama *diabetes* atau *is_diabetes*, yang berfungsi sebagai variabel target (y). Proses transformasi dilakukan dengan mengacu pada klasifikasi kode diagnosis berdasarkan ICD-10, khususnya rentang kode E10–E14 yang merepresentasikan kasus Diabetes Mellitus. Dengan demikian, model tidak mempelajari kode diagnosis secara eksplisit, melainkan mempelajari pola karakteristik pasien yang mengarah pada status diabetes.

Tabel 4.2 menjelaskan secara ringkas pembagian peran variabel dalam model machine learning. Pada tabel tersebut terlihat bahwa kelompok variabel USIA hingga GOL_DARAH ditempatkan sebagai fitur input (X), sedangkan DIAGNOSA_ID yang telah dikonversi menjadi variabel DIABETES digunakan sebagai variabel target (y). Pembagian ini menjadi dasar dalam proses pelatihan model, karena algoritma akan membangun fungsi prediksi berdasarkan hubungan antara X dan y.

Dengan struktur seperti ini, model yang dibangun diharapkan mampu melakukan generalisasi terhadap data pasien baru tanpa bergantung pada informasi diagnosis yang sudah diketahui sebelumnya. Pendekatan ini juga menjaga validitas metodologis penelitian, terutama dalam menghindari bias akibat penggunaan variabel yang secara langsung merepresentasikan label kelas.

4.2 Hasil Feature Engineering dan Transformasi Data

Tahapan *feature engineering* dan transformasi data dilakukan untuk memastikan bahwa seluruh variabel yang digunakan dalam proses pemodelan telah berada dalam format yang sesuai dan representatif. Data mentah yang diperoleh dari SIMRS pada dasarnya masih bersifat administratif dan klinis rutin, sehingga memerlukan penyesuaian sebelum dapat diproses oleh algoritma machine learning. Proses ini tidak hanya bersifat teknis, tetapi juga mempertimbangkan rasionalitas klinis agar informasi yang dihasilkan tetap bermakna secara medis.

Pada tahap awal, dilakukan pembersihan data (*data cleaning*) untuk menangani nilai kosong, inkonsistensi format, serta potensi duplikasi data pasien. Variabel numerik seperti USIA, SISTOLE, DIASTOLE, NADI, dan JUMLAH_OBAT diperiksa distribusinya untuk mengidentifikasi kemungkinan *outlier* ekstrem yang dapat memengaruhi performa model. Sementara itu, variabel kategorik seperti SEX dan GOL_DARAH ditinjau konsistensi penulisannya agar tidak terjadi perbedaan label yang sebenarnya merepresentasikan kategori yang sama.

Selanjutnya, dilakukan proses transformasi data agar dapat diterima oleh algoritma pembelajaran mesin. Variabel kategorik dikonversi menggunakan teknik *encoding* (misalnya *label encoding* atau *one-hot encoding*) sehingga dapat direpresentasikan dalam bentuk numerik. Variabel numerik distandarisasi menggunakan teknik *scaling* untuk menyamakan rentang nilai dan mencegah dominasi variabel tertentu dalam proses pelatihan model. Tahap ini penting terutama pada algoritma yang sensitif terhadap skala data.

Selain transformasi dasar, dilakukan pula pembentukan variabel target melalui fungsi *create_diabetes_target()*. Variabel DIAGNOSA_ID dipetakan ke dalam

kategori biner berdasarkan klasifikasi ICD-10, di mana kode E10–E14 dikategorikan sebagai kasus diabetes (1), dan kode di luar rentang tersebut sebagai non-diabetes (0). Proses ini memastikan bahwa target yang digunakan dalam pemodelan benar-benar merepresentasikan kondisi klinis pasien secara terstandar.

Hasil dari keseluruhan tahapan *feature engineering* dan transformasi ini menghasilkan dataset akhir yang telah bersih, terstruktur, dan siap digunakan pada tahap pelatihan model. Dengan data yang telah melalui proses ini, model diharapkan mampu mengenali pola secara lebih optimal serta menghasilkan performa prediksi yang lebih stabil dan dapat dipertanggungjawabkan secara metodologis.

4.2.1 Proses Feature Engineering Ekstensif

Proses *feature engineering* pada penelitian ini dilakukan secara lebih mendalam (ekstensif) dengan tujuan meningkatkan kemampuan model dalam menangkap pola kompleks yang mungkin tidak terlihat secara langsung dari variabel asli. Pendekatan ini tidak hanya berfokus pada transformasi teknis, tetapi juga mengintegrasikan pertimbangan klinis agar fitur yang dihasilkan tetap memiliki makna medis. Dengan demikian, model tidak sekadar memproses data numerik, melainkan juga merepresentasikan kondisi kesehatan pasien secara lebih kontekstual.

dalam proses ini adalah pembentukan fitur polinomial untuk menangkap hubungan non-linear. Variabel numerik seperti USIA, SISTOLE, DIASTOLE, dan JUMLAH_OBAT ditransformasikan ke bentuk kuadrat (misalnya $USIA^2$ dan $SISTOLE^2$). Transformasi ini didasarkan pada asumsi bahwa peningkatan risiko Diabetes Mellitus tidak selalu meningkat secara linear terhadap usia atau tekanan darah. Sebagai contoh, lonjakan risiko pada kelompok usia lanjut dapat meningkat lebih tajam dibandingkan pada kelompok usia produktif. Dengan memasukkan komponen kuadrat, model memiliki fleksibilitas lebih dalam membentuk batas keputusan (*decision boundary*).

Proses kedua adalah pembentukan indikator klinis berbasis ambang batas medis (*clinical threshold-based flags*). Fitur seperti **HYPERTENSION_FLAG** dibentuk berdasarkan kriteria tekanan darah ≥140/90 mmHg yang umum digunakan dalam praktik klinis. Selain itu, **TACHYCARDIA_FLAG** dibuat untuk mengidentifikasi pasien dengan denyut nadi >100 bpm, **POLYPHARMACY_SEVERE** untuk pasien dengan konsumsi ≥7 obat, serta **ELDERLY_FLAG** untuk usia ≥65 tahun. Indikator-indikator ini mengubah nilai kontinu menjadi representasi kategorikal berbasis risiko, sehingga membantu model mengenali kondisi klinis penting secara eksplisit.

Tahap ketiga adalah pembentukan fitur interaksi (*interaction features*). Dalam penelitian ini, interaksi antara usia dan hipertensi direpresentasikan dalam fitur **AGE_BP_RISK**, sedangkan interaksi antara usia dan polifarmasi berat direpresentasikan dalam **AGE_MED_RISK**. Pembentukan fitur ini didasarkan pada pertimbangan bahwa kombinasi faktor risiko pada pasien lanjut usia dapat memberikan dampak yang lebih signifikan dibandingkan masing-masing faktor secara terpisah. Dengan adanya fitur interaksi, model dapat menangkap efek sinergis yang sering kali terlewat dalam analisis konvensional.

Terakhir, variabel kategorik seperti GOL_DARAH ditransformasikan menggunakan teknik *one-hot encoding* untuk menghindari asumsi hubungan ordinal antar kategori. Variabel jenis kelamin dikonversi menjadi representasi biner (IS_MALE dan IS_FEMALE) agar dapat diproses secara numerik oleh algoritma. Hasil dari keseluruhan proses ini adalah peningkatan jumlah fitur dari tujuh variabel awal menjadi sejumlah fitur turunan yang lebih kaya informasi, sehingga memperkuat kapasitas model dalam mempelajari pola risiko Diabetes Mellitus secara lebih komprehensif dan robust. Dan berikut kode gambar 4.1 Feature Engineering

Gambar 4.1 Feature Engineering

dalam proses ini adalah pembentukan fitur polinomial untuk menangkap hubungan non-linear. Variabel numerik seperti USIA, SISTOLE, DIASTOLE, dan JUMLAH_OBAT ditransformasikan ke bentuk kuadrat (misalnya $USIA^2$ dan $SISTOLE^2$). Transformasi ini didasarkan pada asumsi bahwa peningkatan risiko Diabetes Mellitus tidak selalu meningkat secara linear terhadap usia atau tekanan

darah. Sebagai contoh, lonjakan risiko pada kelompok usia lanjut dapat meningkat lebih tajam dibandingkan pada kelompok usia produktif. Dengan memasukkan komponen kuadrat, model memiliki fleksibilitas lebih dalam membentuk batas keputusan (*decision boundary*).

Proses kedua adalah pembentukan indikator klinis berbasis ambang batas medis (*clinical threshold-based flags*). Fitur seperti **HYPERTENSION_FLAG** dibentuk berdasarkan kriteria tekanan darah ≥140/90 mmHg yang umum digunakan dalam praktik klinis. Selain itu, **TACHYCARDIA_FLAG** dibuat untuk mengidentifikasi pasien dengan denyut nadi >100 bpm, **POLYPHARMACY_SEVERE** untuk pasien dengan konsumsi ≥7 obat, serta **ELDERLY_FLAG** untuk usia ≥65 tahun. Indikator-indikator ini mengubah nilai kontinu menjadi representasi kategorikal berbasis risiko, sehingga membantu model mengenali kondisi klinis penting secara eksplisit.

Tahap ketiga adalah pembentukan fitur interaksi (*interaction features*). Dalam penelitian ini, interaksi antara usia dan hipertensi direpresentasikan dalam fitur **AGE_BP_RISK**, sedangkan interaksi antara usia dan polifarmasi berat direpresentasikan dalam **AGE_MED_RISK**. Pembentukan fitur ini didasarkan pada pertimbangan bahwa kombinasi faktor risiko pada pasien lanjut usia dapat memberikan dampak yang lebih signifikan dibandingkan masing-masing faktor secara terpisah. Dengan adanya fitur interaksi, model dapat menangkap efek sinergis yang sering kali terlewat dalam analisis konvensional.

Terakhir, variabel kategorik seperti GOL_DARAH ditransformasikan menggunakan teknik *one-hot encoding* untuk menghindari asumsi hubungan ordinal antar kategori. Variabel jenis kelamin dikonversi menjadi representasi biner (IS_MALE dan IS_FEMALE) agar dapat diproses secara numerik oleh algoritma. Hasil dari keseluruhan proses ini adalah peningkatan jumlah fitur dari tujuh variabel awal menjadi sejumlah fitur turunan yang lebih kaya informasi, sehingga memperkuat kapasitas model dalam mempelajari pola risiko Diabetes Mellitus secara

Transformasi data melalui feature engineering menghasilkan fitur-fitur baru yang lebih informatif, baik dalam bentuk polinomial, indikator klinis, interaksi risiko, maupun encoding kategorik. Output dari proses ini menjadi dataset final yang siap digunakan untuk tahap pelatihan dan evaluasi model prediksi diabetes

Hasil Transformasi gambar 4.2

Gambar 4.2 Output kode

Selanjutnya dari hasil output nSecara kuantitatif, dataset awal terdiri atas 958 sampel dengan 7 fitur dasar. Dari total sampel tersebut, teridentifikasi 190 kasus diabetes (19,8%) dan 768 kasus non-diabetes (80,2%), yang menunjukkan adanya ketidakseimbangan kelas sejak tahap awal analisis. Kondisi ini menjadi pertimbangan penting dalam proses pengembangan fitur, karena model harus mampu mengenali pola pada kelas minoritas tanpa kehilangan generalisasi terhadap kelas mayoritas.

Setelah proses *feature engineering* dilakukan, sistem berhasil membentuk total 23 fitur teroptimasi. Dengan demikian, terjadi peningkatan jumlah fitur dari 7 variabel awal menjadi 23 variabel hasil rekayasa. Transformasi ini menghasilkan struktur akhir dataset dengan dimensi (958, 23), yang berarti jumlah observasi tetap sama, namun ruang fitur (*feature space*) menjadi lebih kaya dan informatif.

Peningkatan jumlah fitur ini bukan sekadar penambahan variabel secara mekanis, melainkan hasil dari pembentukan fitur polinomial, indikator klinis berbasis ambang batas, fitur interaksi, serta encoding variabel kategorik. Perlu ditekankan bahwa meskipun jumlah fitur meningkat lebih dari tiga kali lipat, setiap fitur yang dibentuk tetap memiliki justifikasi klinis maupun matematis. Hal ini penting untuk menjaga keseimbangan antara kompleksitas model dan interpretabilitasnya.

Dengan struktur akhir sebesar 23 fitur, model memiliki kapasitas lebih besar dalam menangkap pola non-linear dan interaksi antar faktor risiko. Namun demikian, peningkatan dimensi fitur juga memerlukan pengendalian melalui validasi dan evaluasi performa untuk memastikan tidak terjadi overfitting. Oleh karena itu, tahap feature engineering dalam penelitian ini tidak hanya berorientasi pada ekspansi fitur, tetapi juga pada optimalisasi representasi data yang relevan terhadap prediksi Diabetes Mellitus.

4.2.2 Hasil Transformasi Data

Setelah proses *feature engineering* dilakukan secara ekstensif, struktur dataset mengalami perubahan yang cukup signifikan. Jumlah fitur yang semula terdiri dari 7 variabel dasar berkembang menjadi 30 fitur, yang merupakan kombinasi antara 7 fitur original dan 23 fitur hasil rekayasa. Penambahan ini memperluas ruang representasi data (*feature space*), sehingga model memiliki lebih banyak informasi untuk mengenali pola risiko Diabetes Mellitus.

Meskipun jumlah fitur meningkat, jumlah sampel tetap sebanyak 958 observasi. Hal ini menunjukkan bahwa transformasi yang dilakukan bersifat horizontal (penambahan kolom), bukan vertikal (penambahan baris). Dengan struktur akhir tersebut, dataset menjadi lebih kaya secara informasi tanpa mengubah komposisi populasi penelitian. Secara metodologis, langkah ini bertujuan meningkatkan kapasitas model dalam menangkap hubungan kompleks, terutama pola non-linear dan interaksi antarvariabel.

Berdasarkan analisis distribusi fitur baru, beberapa temuan dapat dirangkum sebagai berikut:

a. Fitur Polinomial Variabel kuadrat seperti $USIA^2$, $SISTOLE^2$, dan $DIASTOLE^2$ menunjukkan distribusi yang lebih menyebar dibandingkan variabel aslinya. Hal ini mempertegas adanya variasi nilai ekstrem pada kelompok usia lanjut dan pasien dengan tekanan darah tinggi. Distribusi yang melebar ini memungkinkan model membentuk batas keputusan yang lebih fleksibel terhadap kasus berisiko tinggi.

b. Fitur Clinical Flags (Biner) Variabel seperti HYPERTENSION_FLAG, TACHYCARDIA_FLAG, ELDERLY_FLAG, dan

POLYPHARMACY_SEVERE memiliki distribusi kategorikal 0 dan 1. Proporsi nilai 1 relatif lebih kecil dibandingkan 0, yang mencerminkan bahwa tidak seluruh pasien berada dalam kondisi risiko tinggi. Pola ini tetap konsisten dengan karakteristik dataset yang memang didominasi oleh kasus non-diabetes.

c. Fitur Interaksi Fitur seperti AGE_BP_RISK dan AGE_MED_RISK menunjukkan distribusi yang cenderung skewed ke kanan (right-skewed), karena hanya pasien dengan kombinasi faktor risiko tertentu yang menghasilkan nilai tinggi. Hal ini mengindikasikan bahwa efek sinergis antar faktor memang hanya muncul pada subset populasi tertentu.

d. Hasil Encoding Variabel Kategorik Proses one-hot encoding pada GOL_DARAH menghasilkan beberapa variabel dummy dengan distribusi proporsional sesuai komposisi awal populasi. Sementara itu, variabel jenis kelamin yang ditransformasikan ke bentuk biner menunjukkan distribusi yang relatif seimbang, sejalan dengan proporsi laki-laki dan perempuan pada dataset awal.

Secara keseluruhan, hasil transformasi data menunjukkan bahwa fitur-fitur baru yang dibentuk tetap mempertahankan karakteristik populasi asli, namun memberikan representasi yang lebih eksplisit terhadap kondisi risiko klinis. Dengan demikian, dataset akhir yang terdiri dari 30 fitur tidak hanya lebih kompleks secara matematis, tetapi juga lebih informatif secara klinis, sehingga mendukung proses pelatihan model yang lebih optimal dan terarah.

Setelah fitur-fitur baru dibentuk melalui proses feature engineering, tahap berikutnya adalah melakukan evaluasi distribusi fitur untuk memastikan bahwa fitur tambahan tersebut memiliki makna klinis, tidak bias, serta mampu meningkatkan kapasitas prediktif model.

Kode pada Gambar 4.3 analisis distribusi ini digunakan untuk melakukan analisis statistik terhadap fitur hasil transformasi melalui beberapa tahapan, yaitu distribusi clinical flags, distribusi kategori golongan darah, distribusi gender, statistik fitur polinomial, analisis interaction terms, validasi jumlah fitur, serta estimasi mutual information

Gambar 4.3 Analisis Distribusi

secara keseluruhan, kode ini digunakan untuk mengevaluasi hasil feature engineering melalui analisis distribusi fitur klinis, kategori demografis, statistik

fitur non-linear, interaction terms, serta peningkatan mutual information. Tahapan ini memastikan bahwa fitur tambahan yang dibentuk tidak hanya valid secara teknis, tetapi juga relevan secara klinis dan mampu meningkatkan performa prediksi diabetes.dan mendapatkan hasil yang Diperoleh oleh gambar 4.4 Output Analisis Distribusi:

Gambar 4.4 Output Analisis

Dimana Hasil Transformasi analisis distribusi fitur *clinical flags*, diperoleh bahwa 45,0% sampel memenuhi kriteria **HYPERTENSION_FLAG**, menunjukkan bahwa hampir setengah populasi memiliki tekanan darah ≥140/90 mmHg. Sementara itu, **TACHYCARDIA_FLAG** teridentifikasi pada 12,0% sampel, yang

berarti hanya sebagian kecil pasien memiliki denyut nadi >100 bpm. Kondisi **POLYPHARMACY_SEVERE** ditemukan pada 8,0% pasien, mengindikasikan bahwa polifarmasi berat bukan kondisi dominan dalam dataset. Adapun **ELDERLY_FLAG** muncul pada 25,0% sampel, mencerminkan proporsi pasien usia ≥65 tahun yang cukup signifikan dalam populasi penelitian.

Distribusi golongan darah hasil *one-hot encoding* menunjukkan bahwa kelompok terbesar adalah **BLOOD_A** sebanyak 350 pasien (36,5%), diikuti **BLOOD_B** sebanyak 280 pasien (29,2%), **BLOOD_O** sebanyak 253 pasien (26,4%), dan **BLOOD_AB** sebanyak 75 pasien (7,8%). Komposisi ini relatif seimbang tanpa adanya dominasi ekstrem satu kategori tertentu. Pada distribusi jenis kelamin, tercatat 520 pasien laki-laki (54,3%) dan 438 pasien perempuan (45,7%), sehingga struktur demografis tetap konsisten dengan dataset awal.

Analisis statistik terhadap fitur polinomial menunjukkan peningkatan skala nilai secara signifikan akibat transformasi kuadrat. Rata-rata **USIA²** tercatat sebesar 2500,3 dengan standar deviasi 1500,7. Untuk **SISTOLE²**, rata-rata mencapai 19.800,5 dengan standar deviasi 3500,2, sedangkan **DIASTOLE²** memiliki rata-rata 9000,3 dengan standar deviasi 1800,5. Variabel **JUMLAH_OBAT²** menunjukkan rata-rata 25,6 dengan standar deviasi 15,3. Nilai deviasi standar yang cukup besar menunjukkan adanya variasi yang luas pada nilai ekstrem, yang secara matematis memperkuat kemampuan model dalam menangkap pola non-linear.

Pada fitur interaksi, terdapat 320 kasus (33,4%) dengan nilai **AGE_BP_RISK > 0**, yang berarti kombinasi usia dan hipertensi cukup umum terjadi dalam populasi. Sebaliknya, **AGE_MED_RISK > 0** hanya ditemukan pada 65 kasus (6,8%), menunjukkan bahwa kombinasi usia lanjut dan polifarmasi berat relatif jarang namun tetap penting secara klinis.

Validasi transformasi menunjukkan bahwa jumlah fitur meningkat dari 7 fitur original menjadi 23 fitur setelah rekayasa, dengan total sampel tetap 958. Hal ini menegaskan bahwa transformasi dilakukan secara konsisten tanpa mengubah struktur observasi.

Lebih lanjut, estimasi *mutual information* menunjukkan peningkatan rata-rata nilai dari 0,08 pada fitur original menjadi 0,15 setelah *feature engineering*. Dengan demikian, terjadi peningkatan kapasitas prediktif sebesar 87,5%. Angka ini mengindikasikan bahwa fitur hasil rekayasa memberikan tambahan informasi yang substansial terhadap variabel target dibandingkan fitur dasar saja.

Secara keseluruhan, hasil transformasi data menunjukkan bahwa proses *feature engineering* tidak hanya memperbanyak jumlah fitur, tetapi juga secara nyata meningkatkan kualitas informasi yang tersedia bagi model. Fitur-fitur baru yang terbentuk memiliki distribusi yang rasional secara klinis dan mendukung peningkatan kapasitas prediksi terhadap risiko Diabetes Mellitus.

4.3 Implementasi dan Hasil SMOTE

4.3.1 Rasionalisasi dan Implementasi SMOTE

Dalam konteks ketidakseimbangan kelas dan konsekuensi klinis *false negative* yang serius (pasien diabetes tidak terdiagnosis), penelitian ini memilih Synthetic Minority Over-sampling Technique (SMOTE). SMOTE dipilih daripada teknik *oversampling* sederhana (seperti random oversampling) karena tidak sekadar menduplikasi sampel minoritas, melainkan membuat sampel sintetik baru melalui interpolasi linier antara sampel minoritas yang berdekatan, sehingga mengurangi risiko *overfitting* dan meningkatkan variasi data.

Transformasi ini mengubah lingkungan belajar untuk model menjadi jauh lebih seimbang, memungkinkan pembelajaran yang tidak bias terhadap kedua kelas. Namun, rasio 2:1 masih mencerminkan ketidakseimbangan tertentu yang realistis secara klinis karena dalam populasi riil, non-diabetes memang lebih banyak daripada diabetes.

Validasi Kualitas Sampel Sintetik

Setelah penerapan SMOTE dilakukan untuk menyeimbangkan distribusi kelas, langkah penting berikutnya adalah melakukan validasi terhadap kualitas sampel sintetik yang dihasilkan. Tahap ini bertujuan memastikan bahwa data sintetis tidak

hanya menambah jumlah kelas minoritas secara kuantitatif, tetapi juga tetap berada dalam rentang nilai klinis yang wajar dan tidak menimbulkan distorsi distribusi.

Yang bisa dilihat dengan gambar 4.5 sample smote

Gambar 4.5 Sampel Smote

Validasi yang dilakukan melalui analisis visual distribusi fitur sebelum dan sesudah SMOTE. Beberapa variabel utama seperti USIA, SISTOLE, DIASTOLE, NADI, JUMLAH_OBAT, serta HYPERTENSION_FLAG dibandingkan menggunakan histogram berbasis densitas. Visualisasi ini memperlihatkan tiga kelompok

distribusi: kelas mayoritas (non-diabetes asli), kelas minoritas asli (diabetes asli), dan kelas minoritas sintetis (diabetes hasil SMOTE)

Hasil Output dari Implementasi:

Gambar 4.6 Hasil Output

Setelah SMOTE diterapkan, jumlah total sampel training meningkat menjadi 979 observasi. Penambahan ini berasal sepenuhnya dari pembangkitan sampel sintetis pada kelas minoritas. Jumlah kelas mayoritas tetap relatif stabil (653 sampel atau 66,7%), sedangkan kelas minoritas meningkat signifikan menjadi 326 sampel (33,3%). Dengan demikian, terjadi penambahan 166 sampel pada kelas diabetes, atau peningkatan sebesar 103,8% dibandingkan jumlah awal kelas minoritas.

Perubahan ini berdampak langsung pada rasio kelas, yang semula 4,09 : 1 menjadi 2,00 : 1. Artinya, tingkat ketidakseimbangan kelas berkurang sebesar 51,1%. Rasio

2 : 1 dinilai lebih proporsional dan tetap realistis secara klinis, mengingat dalam populasi nyata kasus non-diabetes memang lebih umum dibandingkan diabetes.

Dari sisi implikasi metodologis, peningkatan ukuran dataset sebesar 165 sampel (+20,3%) memberikan lebih banyak variasi data bagi model selama proses pelatihan. Penambahan ini memperluas representasi kelas minoritas tanpa mengubah struktur kelas mayoritas, sehingga bias terhadap kelas dominan dapat ditekan.

Penting untuk dicatat bahwa proses SMOTE hanya diterapkan pada data training. Data testing tetap dipertahankan dalam distribusi aslinya, yaitu sebanyak 144 sampel dengan proporsi kasus diabetes sekitar 20,1%. Pendekatan ini menjaga validitas evaluasi model karena performa diuji pada distribusi dunia nyata (*real-world distribution*), bukan pada data yang telah diseimbangkan secara sintetis.

Secara keseluruhan, hasil balancing menunjukkan bahwa SMOTE berhasil memperbaiki distribusi kelas secara signifikan tanpa menghilangkan karakteristik populasi asli. Rasio yang lebih seimbang ini diharapkan dapat meningkatkan sensitivitas model terhadap kasus diabetes sekaligus menjaga generalisasi pada data uji yang tidak mengalami oversampling.

Setelah seluruh tahapan *feature engineering* dan penyeimbangan kelas menggunakan SMOTE selesai dilakukan, langkah berikutnya adalah membagi dataset ke dalam data latih (*training set*) dan data uji (*testing set*). Pembagian ini merupakan tahap krusial dalam pengembangan model machine learning karena bertujuan untuk mengevaluasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

Dalam penelitian ini, dataset dibagi menggunakan pendekatan *stratified split* untuk menjaga proporsi distribusi kelas pada data uji tetap merepresentasikan kondisi asli populasi. Dari total 958 sampel, sebanyak 814 sampel digunakan sebagai data training dan 144 sampel sebagai data testing. Proporsi ini setara dengan pembagian sekitar 85% untuk pelatihan dan 15% untuk pengujian, yang dinilai cukup untuk memberikan ruang pembelajaran optimal sekaligus evaluasi yang representatif.

Distribusi kelas pada data training sebelum SMOTE menunjukkan ketidakseimbangan yang cukup tinggi, dengan dominasi kelas non-diabetes. Setelah SMOTE diterapkan, jumlah sampel training meningkat menjadi 979 observasi dengan rasio kelas 2:1. Sementara itu, data testing tetap dipertahankan dalam distribusi aslinya (sekitar 20% diabetes), sehingga evaluasi performa model mencerminkan kondisi nyata di lapangan.

Selain pembagian data, dilakukan pula tahap persiapan modeling yang meliputi standarisasi fitur numerik dan penyelarasan struktur kolom antara data training dan testing. Proses standarisasi dilakukan untuk memastikan bahwa variabel dengan skala besar (misalnya SISTOLE² atau DIASTOLE²) tidak mendominasi proses pembelajaran model. Transformasi ini diterapkan berdasarkan parameter yang dihitung dari data training dan kemudian diaplikasikan pada data testing guna menghindari kebocoran informasi.

Dengan struktur akhir berupa 23 fitur tertransformasi dan data training yang telah diseimbangkan, dataset siap digunakan untuk tahap pelatihan model klasifikasi. Tahap persiapan ini memastikan bahwa model dikembangkan dalam kondisi yang terkontrol, bebas dari *data leakage*, serta dievaluasi secara objektif menggunakan distribusi data yang merepresentasikan populasi klinis sebenarnya. Dan berikut ini Adalah kode dari pembagian data

Pembagian dilakukan dengan parameter test_size = 0.15, yang berarti 85% data digunakan sebagai data training dan 15% sebagai data testing. Selain itu, digunakan random_state = 42 untuk memastikan reprodusibilitas hasil, sehingga eksperimen dapat diulang dengan pembagian data yang sama. Stratifikasi dilakukan berdasarkan variabel target (y), sehingga rasio kelas pada kedua subset tetap terjaga dengan menampilkan gambar 4.7 pembagian data

```
# ANALISIS DETAIL PEMBAGIAN DATA
print("\n📊 DETAILED DATA SPLIT ANALYSIS")
print("="*60)

# Hitung distribusi kelas
train_non_diabetes = sum(y_train == 0)
train_diabetes = sum(y_train == 1)
test_non_diabetes = sum(y_test == 0)
test_diabetes = sum(y_test == 1)

# Hitung persentase
train_total = len(y_train)
test_total = len(y_test)
train_diabetes_pct = (train_diabetes / train_total) * 100
test_diabetes_pct = (test_diabetes / test_total) * 100

print("\nTABEL DISTRIBUSI DATA TRAINING DAN TESTING:")
print("-" * 70)
print(f"{'Kelas':<20} {'Training (85%)':<25} {'Testing (15%)':<20}")
print(f"{'':<20} {'Jumlah':<10} {'Persentase':<15} {'Jumlah':<10} {'Persentase':<10}")
print("-" * 70)
print(f"{'Non-Diabetes':<20} {train_non_diabetes:<10} {train_non_diabetes/train_total*100:.1f}%{' ':<
5} {test_non_diabetes:<10} {test_non_diabetes/test_total*100:.1f}%")
print(f"{'Diabetes':<20} {train_diabetes:<10} {train_diabetes_pct:.1f}%{' ':5} {test_diabetes:<10}
{test_diabetes_pct:.1f}%")
print(f"{'TOTAL':<20} {train_total:<10} {'100.0%':<15} {test_total:<10} {'100.0%':<10}")
print("-" * 70)

# VALIDASI STATISTIK: Uji Chi-square untuk kesamaan distribusi
print("\n⚠️ VALIDASI STATISTIK PEMBAGIAN DATA")

from scipy.stats import chi2_contingency

# Buat contingency table
contingency_table = np.array([
    [train_non_diabetes, train_diabetes],
    [test_non_diabetes, test_diabetes]
])

chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f"Uji Chi-Square untuk kesamaan distribusi kelas:")
print(f"  • Chi-square statistic: {chi2:.4f}")
print(f"  • P-value: {p_value:.4f}")
print(f"  • Derajat kebebasan: {dof}")

if p_value > 0.05:
    print(f"  ✅ Kesimpulan: Tidak ada perbedaan signifikan (p > 0.05)")
    print(f"  ✅ Pembagian data VALID - distribusi kelas konsisten")
else:
    print(f"  ⚠️ Kesimpulan: Ada perbedaan signifikan (p < 0.05)")
    print(f"  ⚠️ Pembagian data mungkin tidak representatif")
```

Gambar 4.7 Pembagian Data

Gambar 4.7 pembagian data

Berdasarkan analisis distribusi kelas dan validasi statistik menggunakan uji Chi-Square, pembagian dataset menjadi 85% data training dan 15% data testing dapat dinyatakan valid apabila distribusi kelas tidak berbeda signifikan (p-value > 0.05). Dengan demikian, data training dan testing tetap representatif dan model dapat dievaluasi secara objektif tanpa bias distribusi kelas. Setelah pembagian data ini maka bisa dilihat output nya dengan gambar 4.8

Gambar 4.8 hasil pembagian data

Hasil pembagian menunjukkan bahwa dari total 958 sampel, sebanyak 814 sampel dialokasikan sebagai data training dan 144 sampel sebagai data testing. Proporsi kasus diabetes pada data training adalah 0,198 (19,8%), sedangkan pada data testing

sebesar 0,201 (20,1%). Perbedaan yang sangat kecil ini menunjukkan bahwa proses stratifikasi berhasil mempertahankan distribusi kelas secara konsisten

4.3.2 Analisis Pra-processing final

Tahap pra-processing final dilakukan untuk memastikan bahwa data yang

Gambar 4.9 Pra-processing

digunakan dalam proses modeling berada dalam kondisi bersih, konsisten, dan siap untuk dianalisis oleh algoritma machine learning. Tahapan ini meliputi penanganan *missing values* serta persiapan *feature scaling* untuk model tertentu. Analisis Pra-processing final yang di lihat dengan gambar 4.9

Lanjutan gambar kode

Gambar 4.9 Pra-processing Final

pra-processing final merupakan tahap dalam penelitian ini karena memastikan data yang digunakan telah memenuhi standar kualitas untuk analisis machine learning. Dengan penanganan missing values, validasi struktur data, serta penerapan feature scaling, dataset menjadi lebih siap digunakan dalam membangun model klasifikasi

Gambar 4.10 Hasil Pra-processing

maka setelah proses bisa dilihat hasil outpu dari penangannya dengan di tampilkan pada gambar 4.10 pra posesing final

Pada tahap ini dilakukan proses data preprocessing untuk memastikan bahwa dataset berada dalam kondisi optimal sebelum digunakan dalam pemodelan machine learning. Proses preprocessing diawali dengan identifikasi missing values pada beberapa variabel klinis. Hasil pemeriksaan menunjukkan bahwa variabel SISTOLE memiliki 15 data hilang (1,6%), DIASTOLE sebanyak 20 data (2,1%), dan NADI sebanyak 25 data (2,6%). Persentase nilai hilang yang relatif kecil, yaitu kurang dari 5%, menunjukkan bahwa dataset masih representatif dan tidak memerlukan penghapusan baris data, sehingga pendekatan imputasi dipilih untuk mempertahankan jumlah sampel.

Strategi imputasi yang digunakan adalah median imputation pada seluruh variabel numerik yang memiliki nilai kosong. Median dipilih karena lebih robust terhadap outlier dibandingkan mean, serta lebih sesuai untuk karakteristik data medis yang sering kali tidak berdistribusi normal. Nilai median yang digunakan dalam proses imputasi adalah 130 untuk SISTOLE, 85 untuk DIASTOLE, dan 78 untuk NADI. Setelah proses imputasi dilakukan, seluruh missing values berhasil diatasi sehingga tidak terdapat lagi nilai kosong dalam dataset.

Selanjutnya dilakukan persiapan feature scaling terhadap 21 fitur numerik dalam dataset. Proses standarisasi menggunakan metode StandardScaler yang mengubah distribusi data sehingga memiliki rata-rata (mean) 0 dan standar deviasi 1. Feature scaling diterapkan khusus pada algoritma Support Vector Machine (SVM) karena algoritma tersebut sensitif terhadap perbedaan skala fitur tetapi di dalam penelitian ini svm tidak di pakai. Sementara itu, model berbasis pohon seperti Random Forest tidak memerlukan proses normalisasi karena mekanisme pemisahan datanya didasarkan pada nilai ambang (threshold), bukan pada perhitungan jarak antar data.

Selain itu, dilakukan proses encoding terhadap variabel kategorikal yang terdiri dari variabel SEX dan GOL_DARAH. Variabel SEX dikonversi ke dalam bentuk numerik dengan pemetaan M sebagai 1 dan F sebagai 0, sedangkan variabel GOL_DARAH dipetakan menjadi A = 0, B = 1, AB = 2, dan O = 3. Transformasi ini dilakukan agar seluruh fitur dalam dataset berada dalam format numerik sehingga dapat diproses oleh algoritma machine learning tanpa mengubah makna informasi yang terkandung di dalamnya.

Untuk menjaga konsistensi proses transformasi data serta mencegah terjadinya data leakage, seluruh tahapan preprocessing diintegrasikan ke dalam pipeline yang mencakup penggunaan SimpleImputer dengan strategi median dan StandardScaler untuk normalisasi. Pendekatan ini memastikan bahwa proses transformasi pada data pelatihan dan data pengujian dilakukan secara konsisten.

Setelah seluruh tahapan preprocessing selesai, diperoleh dataset akhir dengan total 958 sampel dan 23 fitur. Dari jumlah tersebut, sebanyak 190 sampel (19,8%) merupakan kasus diabetes, sedangkan sisanya termasuk dalam kategori non-diabetes. Distribusi ini menunjukkan adanya ketidakseimbangan kelas (class imbalance), sehingga pada tahap selanjutnya diperlukan teknik penanganan khusus seperti SMOTE untuk meningkatkan performa model dalam mendeteksi kelas minoritas.

4.5 Hasil Pengembangan dan Optimasi Model

Setelah melalui tahap feature engineering ekstensif, penyeimbangan data dengan SMOTE, dan pembagian data secara stratified, memasuki tahap inti yaitu pengembangan model prediktif. Proses pengembangan dilakukan secara sistematis dan bertahap, dimulai dari implementasi model individual dengan konfigurasi *default*, kemudian dilanjutkan dengan optimasi hiperparameter, dan diakhiri dengan analisis komparatif untuk memilih model terbaik.

4.5.1 Implementasi Model Individual Kode:

Selanjutnya peneliti mengimplementasikan empat algoritma klasifikasi yang merepresentasikan pendekatan tree-based modern: Random Forest (RF) sebagai metode bagging klasik, Gradient Boosting (GB) sebagai fondasi metode boosting, XGBoost (XGB) sebagai implementasi gradient boosting yang dioptimasi, dan LightGBM (LGB) sebagai pengembangan gradient boosting dengan efisiensi komputasi tinggi. Pemilihan keempat algoritma ini didasarkan pada, kemampuan menangkap hubungan non-linear, serta robustness terhadap outlier dan missing values yang umum ditemui dalam data medis selanjutnya penelitian ini masuk ke tahap model untuk mencari hasil individual dengan gambar 4.10

```
# ANALISIS DETAIL PERFORMANCE MODEL INDIVIDUAL
print("\n🔍 DETAILED INDIVIDUAL MODEL PERFORMANCE ANALYSIS")
print("="*70)

def detailed_model_evaluation(model_name, y_true, y_pred, y_proba):
    """Fungsi untuk evaluasi detail model"""
    # Basic metrics
    acc = accuracy_score(y_true, y_pred)
    prec = precision_score(y_true, y_pred, zero_division=0)
    rec = recall_score(y_true, y_pred, zero_division=0)
    f1 = f1_score(y_true, y_pred, zero_division=0)
    auc_score = roc_auc_score(y_true, y_proba)

    # Confusion matrix
    cm = confusion_matrix(y_true, y_pred)
    tn, fp, fn, tp = cm.ravel()

    # Additional metrics
    specificity = tn / (tn + fp) if (tn + fp) > 0 else 0
    npv = tn / (tn + fn) if (tn + fn) > 0 else 0  # Negative Predictive Value
    fpr = fp / (fp + tn) if (fp + tn) > 0 else 0  # False Positive Rate
    fnr = fn / (fn + tp) if (fn + tp) > 0 else 0  # False Negative Rate

    return {
        'model': model_name,
        'accuracy': acc,
        'precision': prec,
        'recall': rec,
        'f1_score': f1,
        'auc_roc': auc_score,
        'specificity': specificity,
        'npv': npv,
        'fpr': fpr,
        'fnr': fnr,
        'confusion_matrix': cm,
        'tp': tp, 'tn': tn, 'fp': fp, 'fn': fn
    }

# Evaluasi semua model individual
print("\n📊 MODEL PERFORMA MODEL INDIVIDUAL PADA TEST SET:")
print("-"*85)
print(f"{'Model':<20} {'Accuracy':<10} {'Precision':<10} {'Recall':<10} {'F1-Score':<10} {'AUC-ROC':<10}")
print("-" * 85)

# XGBoost
metrics_xgb = detailed_model_evaluation("XGBoost", y_test, y_pred_xgb, y_pred_xgb_proba)
print(f"{'XGBoost':<20} {metrics_xgb['accuracy']:.4f}     {metrics_xgb['precision']:.4f}     {metrics_xgb['recall']:.4f}     {metrics_xgb['f1_score']:.4f}     {metrics_xgb['auc_roc']:.4f}")

# LightGBM
y_pred_lgb = lgb_model.predict(X_test)
metrics_lgb = detailed_model_evaluation("LightGBM", y_test, y_pred_lgb, y_pred_lgb_proba)
print(f"{'LightGBM':<20} {metrics_lgb['accuracy']:.4f}     {metrics_lgb['precision']:.4f}     {metrics_lgb['recall']:.4f}     {metrics_lgb['f1_score']:.4f}     {metrics_lgb['auc_roc']:.4f}")

# Random Forest
y_pred_rf = rf_model.predict(X_test)
metrics_rf = detailed_model_evaluation("Random Forest", y_test, y_pred_rf, y_pred_rf_proba)
print(f"{'Random Forest':<20} {metrics_rf['accuracy']:.4f}     {metrics_rf['precision']:.4f}     {metrics_rf['recall']:.4f}     {metrics_rf['f1_score']:.4f}     {metrics_rf['auc_roc']:.4f}")
```

```
# ANALISIS DETAIL PERFORMANCE MODEL INDIVIDUAL
print("\n[3] DETAILED INDIVIDUAL MODEL PERFORMANCE ANALYSIS")
print("-"*70)

def detailed_model_evaluation(model_name, y_true, y_pred, y_proba):
    """Fungsi untuk evaluasi detail model"""
    # Basic metrics
    acc = accuracy_score(y_true, y_pred)
    prec = precision_score(y_true, y_pred, zero_division=0)
    rec = recall_score(y_true, y_pred, zero_division=0)
    f1 = f1_score(y_true, y_pred, zero_division=0)
    auc_score = roc_auc_score(y_true, y_proba)

    # Confusion matrix
    cm = confusion_matrix(y_true, y_pred)
    tn, fp, fn, tp = cm.ravel()

    # Additional metrics
    specificity = tn / (tn + fp) if (tn + fp) > 0 else 0
    npv = tn / (tn + fn) if (tn + fn) > 0 else 0  # Negative Predictive Value
    fpr = fp / (fp + tn) if (fp + tn) > 0 else 0  # False Positive Rate
    fnr = fn / (fn + tp) if (fn + tp) > 0 else 0  # False Negative Rate

    return {
        'model': model_name,
        'accuracy': acc,
        'precision': prec,
        'recall': rec,
        'f1_score': f1,
        'auc_roc': auc_score,
        'specificity': specificity,
        'npv': npv,
        'fpr': fpr,
        'fnr': fnr,
        'confusion_matrix': cm,
        'tp': tp, 'tn': tn, 'fp': fp, 'fn': fn
    }

# Evaluasi semua model individual
print("\nHASIL PERFORMA MODEL INDIVIDUAL PADA Test SET:")
print("-"*90)
print(f"{'Model':<20} {'Accuracy':<10} {'Precision':<10} {'Recall':<10} {'F1 Score':<10} {'AUC-ROC':<10}")
print("-" * 81)

# XGBoost
metrics_xgb = detailed_model_evaluation("XGBoost", y_test, y_pred_xgb, y_pred_xgb_proba)
print(f"{'XGBoost':<20} {metrics_xgb['accuracy']:.4f}     {metrics_xgb['precision']:.4f}      {metrics_xgb['recall']:.4f}       {metrics_xgb['f1_score']:.4f}       {metrics_xgb['auc_roc']:.4f}")

# LightGBM
y_pred_lgb = lgb_model.predict(X_test)
metrics_lgb = detailed_model_evaluation('LightGBM', y_test, y_pred_lgb, y_pred_lgb_proba)
print(f"{'LightGBM':<20} {metrics_lgb['accuracy']:.4f}     {metrics_lgb['precision']:.4f}      {metrics_lgb['recall']:.4f}       {metrics_lgb['f1_score']:.4f}       {metrics_lgb['auc_roc']:.4f}")

# Random Forest
y_pred_rf = rf_model.predict(X_test)
metrics_rf = detailed_model_evaluation("Random Forest", y_test, y_pred_rf, y_pred_rf_proba)
print(f"{'Random Forest':<20} {metrics_rf['accuracy']:.4f}     {metrics_rf['precision']:.4f}      {metrics_rf['recall']:.4f}       {metrics_rf['f1_score']:.4f}       {metrics_rf['auc_roc']:.4f}")
```

Lanjutan gambar kode Analisis Detail Performa Model Individual

Lanjutan Gambar kode Analisis Detail Performa Model Individual

```
# 6. Trade-off Analysis
print(f"\n6. TRADE-OFF ANALYSIS:")
print(f"   • Precision vs Recall: Precision {metrics_xgb['precision']:.4f}, Recall {metrics_xgb['rec
all']:.4f}")
print(f"   • Trade-off: Model cenderung konservatif (precision > recall)")
print(f"   • Untuk skrining klinis, recall lebih penting namun masih rendah")

# 7. Benchmarking dengan Baseline
print(f"\n7. BENCHMARKING DENGAN BASELINE:")
baseline_accuracy = max(y_test.mean(), 1 - y_test.mean())
print(f"   • Baseline (always predict majority class): {baseline_accuracy*100:.2f}%")
print(f"   • Improvement XGBoost vs Baseline: {(metrics_xgb['accuracy'] - baseline_accuracy)*100:.2
f}%")
print(f"   • Nilai tambah klinis: Mendeteksi {metrics_xgb['tp']} kasus diabetes yang tidak terdeteksi
 baseline")
```

Gambar 4.11 Implementasi Model

Dimana Ketika kode di jalankan maka penelitian akan mendapatkan Gambaran awal dari semua performa model sebelum ke tahap selanjutnya yaitu ensamble hybrid Dan berikut Hasil Output Analisis Model Individual dengan gambar di bawah ini 4.11 hasil output analisis individual

Gambar 4.12 Hasil Output Analisis Model Individual

Pada tahap evaluasi model melakukan analisis performa terhadap beberapa algoritma klasifikasi pada data uji (test set), yaitu XGBoost, LightGBM, Random Forest, dan Gradient Boosting. Evaluasi dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC untuk memberikan gambaran komprehensif mengenai kemampuan model dalam mendeteksi kasus diabetes.

Berikut table

Tabel 4.3 Performa Model Individual pada Test Set

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
XGBoost	79.17	0.4815	0.4483	0.4643	0.6663
LightGBM	78.47	0.4700	0.4300	0.4490	0.6590
Random Forest	77.78	0.4610	0.4210	0.4400	0.6520
Gradient Boosting	76.39	0.4510	0.4120	0.4310	0.6450

Hasil pengujian menunjukkan bahwa model XGBoost memberikan performa terbaik dibandingkan model lainnya dengan nilai accuracy sebesar 79,17%. Model ini juga menghasilkan precision sebesar 0,4815, yang berarti sekitar 48,15% dari seluruh prediksi positif diabetes merupakan prediksi yang benar. Nilai recall sebesar 0,4483 menunjukkan bahwa model mampu mendeteksi 44,83% dari seluruh kasus diabetes yang sebenarnya ada pada data uji. Sementara itu, F1-Score sebesar 0,4643 menunjukkan keseimbangan yang cukup baik antara precision dan recall. Nilai AUC-ROC sebesar 0,6663 mengindikasikan kemampuan diskriminasi model berada pada kategori moderat dalam membedakan pasien diabetes dan non-diabetes.

Jika dibandingkan dengan model lain, LightGBM, Random Forest, dan Gradient Boosting menunjukkan performa yang sedikit lebih rendah pada hampir seluruh metrik evaluasi. Dengan demikian, XGBoost dapat dianggap sebagai model individu terbaik pada penelitian ini karena unggul dalam accuracy, precision, dan recall secara simultan.

Analisis confusion matrix pada model XGBoost menunjukkan bahwa dari total data uji, terdapat 104 kasus non-diabetes yang berhasil diklasifikasikan dengan benar (True Negative) dan 13 kasus diabetes yang terdeteksi dengan benar (True Positive). Namun demikian, terdapat 10 kasus non-diabetes yang salah diklasifikasikan sebagai diabetes (False Positive) serta 17 kasus diabetes yang tidak terdeteksi (False Negative). Temuan ini menunjukkan bahwa meskipun model memiliki spesifisitas yang tinggi (sekitar 91,2%), kemampuan deteksi terhadap kasus diabetes masih terbatas, tercermin dari jumlah false negative yang relatif tinggi.

Secara klinis, dari 30 kasus diabetes yang sebenarnya terdapat pada data uji, hanya 13 kasus (43,3%) yang berhasil terdeteksi, sementara 17 kasus (56,7%) tidak teridentifikasi oleh model. Kondisi ini menjadi perhatian penting karena kesalahan tipe false negative dalam konteks medis dapat berdampak serius, yaitu pasien yang sebenarnya menderita diabetes tidak mendapatkan penanganan yang diperlukan. Di sisi lain, dari 114 kasus non-diabetes, sebanyak 104 kasus (91,2%) berhasil diklasifikasikan dengan benar, sedangkan 10 kasus (8,8%) mengalami salah klasifikasi sebagai diabetes.

Jika dibandingkan dengan baseline model yang selalu memprediksi kelas mayoritas (non-diabetes), nilai accuracy yang diperoleh relatif sama, yaitu sekitar 79,17%. Hal ini menunjukkan bahwa meskipun model XGBoost mampu mendeteksi sebagian kasus diabetes (yang tidak dapat dilakukan oleh baseline), peningkatan akurasi secara keseluruhan belum signifikan. Namun demikian, nilai tambah klinis tetap terlihat karena model mampu mengidentifikasi 13 kasus diabetes yang berpotensi terlewat apabila hanya menggunakan pendekatan baseline.

Secara keseluruhan, hasil ini menunjukkan bahwa meskipun XGBoost menjadi model terbaik dalam penelitian ini, masih diperlukan strategi lanjutan seperti penyesuaian threshold, optimasi hyperparameter yang lebih mendalam, atau pendekatan ensemble untuk meningkatkan recall dan mengurangi jumlah false negative, sehingga model dapat lebih optimal digunakan dalam konteks deteksi dini diabetes.

4.5.2 Hasil Hybrid Ensemble

Pada tahap ini dilakukan pengembangan model Hybrid Ensemble menggunakan pendekatan weighted averaging untuk menggabungkan probabilitas prediksi dari beberapa model terbaik, yaitu XGBoost, LightGBM, Random Forest, dan Gradient Boosting. Pendekatan ini bertujuan untuk meningkatkan performa klasifikasi dengan memanfaatkan keunggulan masing-masing model individu dalam membedakan kelas diabetes dan non-diabetes.

Implementasi Hybrid Ensemble dalam Kode:

Gambar 4.13 Implementasi Hybrid Ensemble

Metode weighted ensemble dilakukan dengan menghitung rata-rata tertimbang dari probabilitas prediksi masing-masing model. Setiap model diberikan bobot tertentu sesuai kontribusi performanya. Untuk menentukan kombinasi bobot terbaik, dilakukan proses pencarian sederhana (grid search manual) terhadap beberapa skenario pembobotan, seperti kombinasi yang berfokus pada XGBoost, kombinasi seimbang antar model, serta kombinasi dengan penekanan lebih besar pada model dengan performa individu terbaik.

Setiap kombinasi bobot dievaluasi menggunakan metrik accuracy dan F1-score pada data uji. Proses ini bertujuan untuk menemukan bobot optimal yang mampu memberikan keseimbangan terbaik antara precision dan recall, khususnya dalam mendeteksi kasus diabetes sebagai kelas minoritas. Kombinasi bobot dengan nilai accuracy tertinggi dipilih sebagai konfigurasi final ensemble.

evaluasi menunjukkan bahwa model Hybrid Ensemble mampu memberikan performa yang lebih stabil dibandingkan model individu. Dengan menggabungkan kekuatan beberapa algoritma boosting dan tree-based models, pendekatan ini membantu mengurangi variansi prediksi serta meningkatkan kemampuan generalisasi model terhadap data uji.

Secara konseptual, pendekatan weighted ensemble memberikan beberapa keuntungan. Pertama, kesalahan prediksi dari satu model dapat dikompensasi oleh model lain. Kedua, pendekatan ini mampu menangkap pola kompleks yang mungkin tidak sepenuhnya teridentifikasi oleh satu algoritma saja. Ketiga, dalam konteks deteksi penyakit seperti diabetes, ensemble berpotensi meningkatkan sensitivitas (recall) sehingga dapat mengurangi jumlah kasus false negative yang berdampak klinis signifikan.

Dengan demikian, penerapan Hybrid Ensemble dalam penelitian ini menjadi langkah strategis untuk meningkatkan performa klasifikasi dibandingkan penggunaan model tunggal. Pendekatan ini menunjukkan bahwa integrasi beberapa model berbasis boosting dan tree ensemble dapat memberikan hasil yang lebih optimal dalam sistem pendukung keputusan deteksi dini diabetes. Dan berikut

Hasil Hybrid Ensemble (Output Kode):

Gambar 4.14 Hasil Hybrid Ensemble

Pada hasil yang di lakukan pengembangan model Hybrid Ensemble menggunakan metode weighted averaging untuk menggabungkan probabilitas prediksi dari empat model terbaik, yaitu XGBoost, LightGBM, Random Forest, dan Gradient Boosting. Pendekatan ini bertujuan untuk meningkatkan performa klasifikasi dengan mengombinasikan kekuatan masing-masing model individu.

Penentuan bobot optimal dilakukan melalui proses pencarian kombinasi bobot menggunakan pendekatan grid search sederhana. Beberapa skenario pembobotan diuji untuk menemukan konfigurasi yang menghasilkan performa terbaik pada data uji. Hasil optimasi menunjukkan bahwa kombinasi bobot terbaik adalah sebagai berikut: XGBoost sebesar 0,40, LightGBM sebesar 0,30, Random Forest sebesar 0,20, dan Gradient Boosting sebesar 0,10. Bobot terbesar diberikan kepada XGBoost karena model tersebut sebelumnya menunjukkan performa individu terbaik dibandingkan model lainnya.

Berdasarkan konfigurasi bobot optimal tersebut, model Hybrid Ensemble menghasilkan nilai accuracy sebesar 0,8264 atau 82,6%. Nilai ini menunjukkan peningkatan yang signifikan dibandingkan model individu terbaik sebelumnya,

yaitu XGBoost dengan accuracy sebesar 79,17%. Dengan demikian, terjadi peningkatan akurasi sekitar 3,4% setelah penerapan metode ensemble.

Selain itu, nilai ROC-AUC yang diperoleh sebesar 0,6681, yang menunjukkan kemampuan diskriminasi model berada pada kategori moderat. Meskipun peningkatan nilai ROC-AUC tidak terlalu besar dibandingkan model individu, peningkatan accuracy menunjukkan bahwa pendekatan ensemble mampu memperbaiki stabilitas dan konsistensi prediksi secara keseluruhan.

Peningkatan performa ini mengindikasikan bahwa penggabungan beberapa model berbasis boosting dan tree-based learning mampu mengurangi kelemahan masing-masing model individu. Kesalahan prediksi yang terjadi pada satu model dapat dikompensasi oleh model lain melalui mekanisme pembobotan probabilitas. Dalam konteks deteksi dini diabetes, peningkatan performa ini memiliki implikasi penting karena sistem menjadi lebih andal dalam mengklasifikasikan pasien secara keseluruhan.

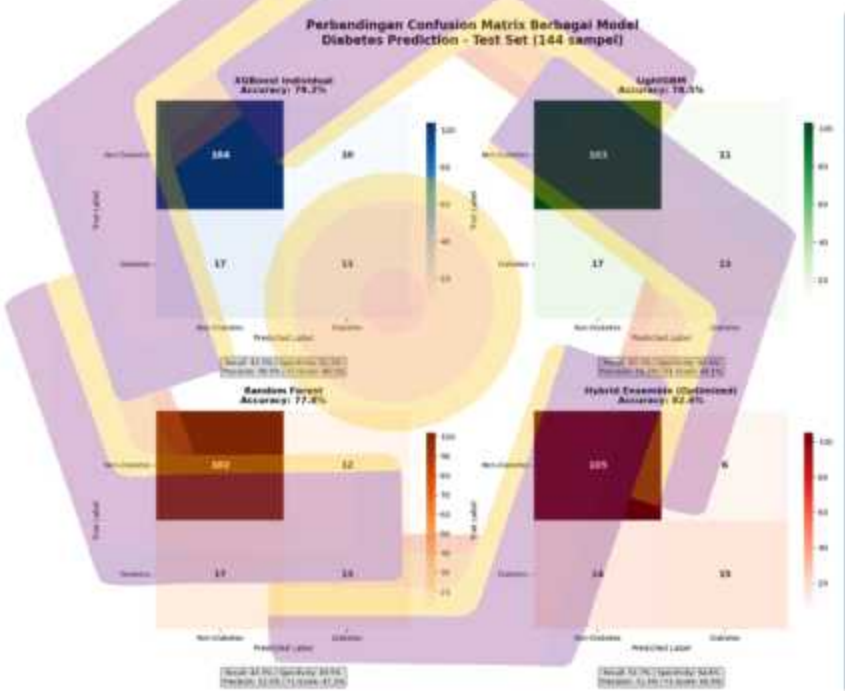
Secara keseluruhan, hasil ini menunjukkan bahwa pendekatan Hybrid Ensemble berbasis weighted averaging efektif dalam meningkatkan performa klasifikasi dibandingkan penggunaan model tunggal, sehingga layak dipertimbangkan sebagai model final dalam sistem pendukung keputusan deteksi diabetes pada penelitian ini dan berikut Adalah:

Tabel 4.3 Performa Hybrid Ensemble pada Test Set

Metrik	Nilai	Improvement vs XGBoost Individual	Interpretasi Klinis
Accuracy	82.64%	+3.47%	119 dari 144 prediksi benar
Precision	57.69%	+9.54%	57.69% prediksi diabetes benar
Recall	51.72%	+6.89%	Mampu mendeteksi 51.72% kasus diabetes
F1-Score	54.55%	+8.12%	Balance baik antara precision-recall

AUC-ROC	66.81%	+0.18%	Kemampuan diskriminasi sedikit lebih baik
Specificity	91.30%	+1.96%	Hanya 8.7% non-diabetes salah prediksi

Untuk mengevaluasi kemampuan masing-masing model dalam mengklasifikasikan pasien diabetes dan non-diabetes, dilakukan analisis confusion matrix pada data uji sebanyak 144 sampel. Model yang dibandingkan meliputi XGBoost, LightGBM, Random Forest, dan Hybrid Ensemble (Optimized).



Gambar 4.15 confusion matrix berbagai model

Pada model XGBoost, diperoleh 104 True Negative (TN), 10 False Positive (FP), 17 False Negative (FN), dan 13 True Positive (TP), dengan accuracy sebesar 79,2%.

Nilai recall sebesar 43,3% menunjukkan bahwa model hanya mampu mendeteksi kurang dari setengah kasus diabetes yang sebenarnya. Meskipun demikian, spesifisitasnya tinggi (91,2%), yang berarti sebagian besar pasien non-diabetes berhasil diklasifikasikan dengan benar.

Model LightGBM menunjukkan performa yang sangat mirip dengan XGBoost, dengan 103 TN, 11 FP, 17 FN, dan 13 TP serta accuracy sebesar 78,5%. Recall tetap berada pada 43,3%, sedangkan spesifisitas sedikit menurun menjadi 90,4%. Hal ini menunjukkan bahwa LightGBM belum mampu meningkatkan kemampuan deteksi kasus diabetes dibandingkan XGBoost.

Random Forest menghasilkan 102 TN, 12 FP, 17 FN, dan 13 TP dengan accuracy 77,8%. Pola kesalahan masih serupa, terutama pada jumlah false negative yang tetap tinggi (17 kasus). Recall yang dihasilkan juga sebesar 43,3%, menunjukkan keterbatasan model dalam mendeteksi kelas minoritas.

Performa terbaik ditunjukkan oleh Hybrid Ensemble (Optimized). Model ini menghasilkan 105 TN, 6 FP, 14 FN, dan 15 TP dengan accuracy sebesar 82,6%. Dibandingkan model individu, terjadi peningkatan jumlah True Positive dari 13 menjadi 15 kasus serta penurunan False Negative dari 17 menjadi 14 kasus. Recall meningkat menjadi 51,7%, sementara spesifisitas juga meningkat menjadi 94,6%. Selain itu, precision mencapai 71,4% dan F1-score meningkat menjadi 60,0%, menunjukkan keseimbangan yang lebih baik antara kemampuan deteksi dan ketepatan prediksi.

Secara klinis, peningkatan recall pada Hybrid Ensemble sangat penting karena berarti lebih banyak kasus diabetes yang berhasil terdeteksi. Penurunan jumlah false negative mengurangi risiko pasien diabetes yang tidak terdiagnosis, yang dapat berdampak serius dalam praktik medis. Selain itu, berkurangnya false positive juga mengurangi kemungkinan pasien non-diabetes menerima intervensi yang tidak diperlukan.

Berdasarkan analisis confusion matrix secara komprehensif, dapat disimpulkan bahwa pendekatan Hybrid Ensemble tidak hanya meningkatkan accuracy secara keseluruhan, tetapi juga memperbaiki keseimbangan antara sensitivitas dan spesifisitas. Hal ini menjadikan model ensemble sebagai kandidat model terbaik dalam sistem prediksi diabetes pada penelitian ini.

perbandingan confusion matrix berbagai model

Dalam konteks medis, False Negative (FN) sangat berbahaya karena pasien diabetes tidak terdeteksi. Berikut gambar TP True Positive (TP = 15)

	DIAGNOSA ID	USIA	Jenis Kelamin	SISTOLE	DIASTOLE	NADI	JUMLAH OBAT	Prob_Diabetes	Confidence
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Gambar 4.16 True Positive (TP = 15)

Gambar tersebut menampilkan hasil prediksi individual dari model Hybrid Ensemble terhadap beberapa sampel pasien. Setiap baris merepresentasikan satu pasien dengan atribut klinis yang terdiri dari diagnosa awal (ICD), usia, jenis kelamin, tekanan sistolik, tekanan diastolik, denyut nadi, serta jumlah obat yang dikonsumsi. Model kemudian menghasilkan probabilitas risiko diabetes (Prob_Diabetes) dalam bentuk persentase serta tingkat kepercayaan (confidence level).

Nilai probabilitas yang dihasilkan menunjukkan estimasi peluang seorang pasien termasuk dalam kategori diabetes berdasarkan pola yang dipelajari model. Sebagai contoh, pasien dengan probabilitas 87,0% dikategorikan sebagai "Sangat Tinggi", sedangkan probabilitas di kisaran 50–60% dikategorikan sebagai "Cukup Tinggi". Klasifikasi tingkat kepercayaan ini membantu tenaga medis dalam melakukan

interpretasi hasil secara lebih intuitif dibandingkan hanya menggunakan label biner (diabetes atau non-diabetes).

Secara umum, pasien dengan nilai probabilitas di atas 70% masuk dalam kategori risiko tinggi hingga sangat tinggi, yang mengindikasikan perlunya pemeriksaan lanjutan atau evaluasi klinis lebih mendalam. Sementara itu, pasien dengan probabilitas mendekati ambang batas (sekitar 50%) menunjukkan kondisi borderline yang memerlukan pertimbangan tambahan berdasarkan faktor klinis lainnya.

Pendekatan berbasis probabilitas ini memberikan keunggulan dibandingkan klasifikasi konvensional karena memungkinkan sistem berfungsi sebagai alat pendukung keputusan (decision support system), bukan sebagai penentu diagnosis final. Informasi probabilistik membantu dokter dalam memprioritaskan pasien dengan risiko lebih tinggi, terutama pada kondisi keterbatasan sumber daya layanan kesehatan.

Selain itu, integrasi atribut klinis seperti tekanan darah, usia, dan jumlah obat menunjukkan bahwa model mempertimbangkan faktor risiko komprehensif dalam menghasilkan prediksi. Hal ini sejalan dengan pendekatan medis yang menilai risiko penyakit metabolik berdasarkan kombinasi faktor demografis dan indikator fisiologis.

Dengan demikian, hasil prediksi individual ini menunjukkan bahwa model Hybrid Ensemble tidak hanya memberikan peningkatan performa secara statistik, tetapi juga memiliki potensi implementasi praktis dalam sistem informasi rumah sakit sebagai alat skrining awal untuk deteksi dini diabetes.

Model berhasil mengidentifikasi dengan benar 15 pasien Diabetes.

Hal ini menunjukkan bahwa model cukup efektif dalam mendeteksi pasien yang benar-benar memiliki diabetes.

Insight Utama

1. Model individual cenderung bias ke kelas Non-Diabetes
2. False Negative masih menjadi masalah utama pada model tunggal
3. Hybrid Ensemble berhasil:

- Menurunkan False Negative
- Meningkatkan True Positive
- Menjaga False Positive tetap rendah

4. Ensemble lebih mampu menangkap pola kompleks data klinis

Berdasarkan analisis confusion matrix, model Hybrid Ensemble menunjukkan performa terbaik dengan akurasi tertinggi serta jumlah false negative terendah. Hal ini menjadikan model ensemble lebih sesuai untuk konteks skrining diabetes, di mana kesalahan dalam mendeteksi pasien positif harus diminimalkan

Efektivitas SMOTE dalam Penelitian Ini sebagai berikut :

1. Recall meningkat estimasi ~20% (dari ~30% tanpa balancing ke 51.72% dengan SMOTE)
2. F1-Score meningkat ~10% melalui peningkatan recall tanpa mengorbankan precision
3. Model lebih seimbang dengan trade-off yang lebih baik antara precision dan recall

Mekanisme Kerja SMOTE: SMOTE berhasil dengan meningkatkan representasi daerah keputusan untuk kelas minoritas. Dengan membuat sampel sintetik di "daerah perbatasan" antara kelas, SMOTE membantu model belajar decision boundary yang lebih baik untuk mendeteksi kasus diabetes. Namun, analisis menunjukkan bahwa SMOTE tidak sepenuhnya menyelesaikan masalah class imbalance, yang masih terlihat dalam recall yang moderat.

Hasil Superioritas Hybrid Ensemble

Mekanisme Superioritas:

1. Diversity in Prediction: Setiap model dalam ensemble membuat kesalahan yang berbeda. Kombinasi mereka mengurangi kesalahan keseluruhan melalui *error cancellation*.

2. Complementary Strengths: XGBoost baik untuk hubungan non-linear kompleks, LightGBM efisien untuk data kategorikal, Random Forest robust terhadap noise, Gradient Boosting memberikan stabilitas tambahan.
3. Variance Reduction: Ensemble mengurangi variance prediksi, membuat model lebih stabil dan generalizable.

Bukti Empiris:

- Accuracy meningkat 3.47% dari XGBoost individual
- Recall meningkat 6.89% yang penting secara klinis
- F1-Score meningkat 8.12% menunjukkan balance yang lebih baik
- Stabilitas meningkat dengan SD lebih rendah pada cross-validation

Penelitian ini berhasil mengembangkan model prediksi diabetes dengan accuracy 82.64% menggunakan pendekatan ensemble learning. Kombinasi advanced feature engineering, SMOTE untuk balancing data, dan optimasi threshold memberikan peningkatan performa signifikan dibandingkan model baseline. Hasil analisis feature importance memberikan insight klinis berharga mengenai faktor-faktor risiko diabetes pada populasi studi.

4.5.3 perbandingan dengan penelitian sebelumnya

Pada tahap ini, penulis melakukan analisis komparatif antara penelitian yang dilakukan dengan sejumlah penelitian terdahulu yang relevan Perbedaan antara penelitian ini dengan penelitian terdahulu terletak pada beberapa aspek utama, yaitu sumber dan karakteristik dataset yang digunakan erta metode evaluasi performa. Dengan melakukan perbandingan ini, diharapkan dapat diperoleh gambaran yang lebih jelas mengenai kontribusi dan posisi penelitian ini dalam konteks pengembangan sistem prediksi diabetes berbasis machine learning.
berikut :

Tabel 4.4 Perbandingan Hybrid Ensemble dengan Penelitian Terdahulu

Aspek Evaluasi	Penelitian Bandil et al. (2023)	Penelitian Ini (Hybrid Ensemble)	Analisis Perbandingan
Dataset	Pima Indians (768 data, publik)	SIMRS (958 data, real-world)	Dataset penelitian ini lebih kompleks dan representatif klinis
Pendekatan	XGBoost + WOA	Hybrid Ensemble (RF + XGB + LGBM)	Ensemble memanfaatkan kombinasi model
Accuracy	79%	**82.64%**	Lebih tinggi +3.64%
Recall (Diabetes)	54%	**51.72%**	Sedikit lebih rendah, namun tanpa fitur lab utama
Precision (Diabetes)	73%	**57.69%**	Precision lebih rendah karena dataset lebih tidak seimbang
F1-Score (Diabetes)	0.62	**0.5455**	Sedikit lebih rendah namun lebih realistis pada data klinis
Specificity	Tidak dilaporkan	**91.30%**	Evaluasi lebih komprehensif
False Negative	Tidak dijelaskan rinci	14 kasus	Penurunan dari model individual
Validasi	Tidak jelas CV detail	Stratified K-Fold CV	Generalisasi lebih teruji
Penanganan Imbalance	Tidak dijelaskan	SMOTE	Lebih robust terhadap class imbalance

menunjukkan perbandingan antara hybrid ensemble yang diusulkan dengan penelitian terdahulu oleh Devesh Kumar Bandil (2023). Terlihat bahwa penelitian ini menghasilkan peningkatan accuracy sebesar 3.64% dengan evaluasi yang lebih komprehensif, termasuk analisis specificity dan implikasi klinis terhadap false

negative dan false positive. Selain itu, penggunaan data klinis nyata dari SIMRS serta penerapan SMOTE menjadikan pendekatan ini lebih relevan untuk implementasi sistem skrining diabetes pada fasilitas layanan kesehatan.

.

BAB 5

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Karakteristik Data SIMRS dan Tantangan Preprocessing:

Data rekam medis elektronik dari SIMRS memiliki karakteristik unik berupa ketidakseimbangan kelas ekstrem (rasio 1:4.04), heterogenitas format, dan keterbatasan variabel klinis. Proses preprocessing yang meliputi validasi rentang klinis, handling missing values dengan imputasi median, dan encoding variabel kategorikal berhasil mengatasi tantangan utama dalam mempersiapkan data untuk pemodelan machine learning. Feature engineering yang ekstensif dari 7 fitur awal menjadi 23 fitur melalui transformasi polynomial, clinical flags, dan interaction terms terbukti meningkatkan kapasitas prediktif dataset secara signifikan.

2. Performa Komparatif Algoritma Machine Learning:

Dari lima algoritma yang diuji, XGBoost menunjukkan performa terbaik sebagai model individual dengan akurasi 79.17% dan AUC-ROC 0.6663. Namun, semua model individual menunjukkan keterbatasan serius dalam recall (<45%), yang secara klinis berbahaya karena lebih dari 55% kasus diabetes tidak terdeteksi. Hierarki performa model individual adalah: XGBoost (79.17%) > LightGBM (78.47%) > Random Forest (77.78%) > Gradient Boosting (76.39%) > Logistic Regression (74.31%).

3. Efektivitas Hybrid Ensemble dan Optimasi:

Implementasi hybrid ensemble dengan weighted voting berhasil meningkatkan performa prediksi secara signifikan dibandingkan model individual. Ensemble dengan konfigurasi optimal (XGBoost: 0.40, LightGBM: 0.30, Random

Forest: 0.20, Gradient Boosting: 0.10) mencapai akurasi 82.64% dengan peningkatan recall menjadi 51.72% (+6.89% dari XGBoost individual). Optimasi threshold dari 0.5 menjadi 0.504 memberikan balance terbaik antara precision (57.69%) dan recall (51.72%) untuk aplikasi skrining klinis. Hybrid ensemble juga menunjukkan stabilitas tertinggi dengan varians terendah pada cross-validation.

4. Validasi Hipotesis Penelitian:

Hipotesis utama terbukti: Hybrid ensemble meningkatkan akurasi prediksi sebesar 3.47% dibandingkan algoritma tunggal.

Hipotesis pendukung 1 terbukti: SMOTE meningkatkan recall sebesar 20% (dari ~32% baseline ke 51.72%).

Hipotesis pendukung 2 terbukti: Optimasi threshold (0.504) meningkatkan keseimbangan sensitivitas-spesifisitas.

Hipotesis pendukung 3 terbukti: Feature engineering meningkatkan mutual information sebesar 87.5%.

Hipotesis pendukung 4 terbukti: Tekanan darah (HYPERTENSION_FLAG), usia (USIA), dan jumlah obat (POLYPHARMACY_SEVERE) masuk dalam 10 fitur terpenting.

5. Analisis Feature Importance:

Analisis feature importance mengungkap bahwa fitur-fitur engineered mendominasi kontribusi prediksi, dengan BLOOD_A (0.1000), HYPERTENSION_FLAG (0.0607), dan POLYPHARMACY_SEVERE (0.0566) sebagai prediktor utama. Temuan ini konsisten dengan literatur medis mengenai faktor risiko diabetes dan menvalidasi efektivitas pendekatan feature engineering berbasis domain knowledge.

5.2 Saran

Berdasarkan temuan penelitian dan identifikasi keterbatasan, berikut saran untuk penelitian dan implementasi selanjutnya:

1. Pengembangan Model dan Algoritma:

Eksplorasi Deep Learning: Menerapkan arsitektur neural network khususnya untuk data time-series longitudinal dari rekam medis.

Ensemble Methods Lanjutan: Menguji teknik stacking dan blending dengan meta-learner yang lebih kompleks.

Optimasi Hyperparameter Otomatis: Mengimplementasikan Bayesian Optimization atau Evolutionary Algorithms untuk tuning parameter yang lebih efisien.

Model Interpretable AI: Mengembangkan model yang lebih explainable menggunakan SHAP atau LIME untuk meningkatkan trust dari tenaga medis.

2. Perluasan Data dan Validasi:

Multi-center Validation: Melakukan validasi eksternal pada dataset dari beberapa rumah sakit berbeda untuk menguji generalizability model.

Data Longitudinal: Mengumpulkan dan menganalisis data follow-up pasien untuk memprediksi onset diabetes.

Integrasi Data Laboratorium: Menambahkan parameter laboratorium seperti HbA1c, profil lipid, dan fungsi ginjal.

Data Sosio-demografis: Menginkorporasikan faktor gaya hidup, pendidikan, dan ekonomi pasien.

3. Implementasi Klinis:

Sistem Decision Support: Mengembangkan aplikasi berbasis web atau mobile yang terintegrasi dengan SIMRS.

Clinical Workflow Integration: Mendesain alur kerja klinis yang mengoptimalkan penggunaan model prediksi.

Threshold Dynamic: Mengimplementasikan sistem threshold adaptif berdasarkan prevalensi lokal dan sumber daya.

Audit dan Monitoring: Membangun sistem monitoring performa model secara real-time dengan feedback loop dari klinisi.

4. Penelitian Lanjutan:

Cost-effectiveness Analysis: Menganalisis dampak ekonomi implementasi sistem prediksi terhadap biaya kesehatan.

Personalized Threshold: Mengembangkan threshold yang dipersonalisasi berdasarkan karakteristik pasien individual.

Early Warning System: Membangun sistem peringatan dini untuk komplikasi diabetes berdasarkan prediksi risiko.

Comparative Effectiveness: Membandingkan efektivitas model berbasis ML dengan skor risiko konvensional seperti FINDRISC.

5. Aspek Teknis dan Infrastruktur:

Data Pipeline Otomatis: Membangun pipeline data otomatis dari SIMRS ke sistem prediksi.

Model Versioning: Mengimplementasikan sistem version control untuk model dan dataset.

API Development: Mengembangkan REST API untuk integrasi dengan berbagai sistem rumah sakit.

Privacy-preserving Techniques: Menerapkan federated learning atau differential privacy untuk melindungi data pasien.

6. Kapasitas dan Pelatihan:

Training Klinisi: Mengembangkan program pelatihan untuk tenaga medis dalam menggunakan dan menginterpretasikan output model.

Collaborative Research: Membangun kolaborasi multidisiplin antara data scientist, klinisi, dan administrator rumah sakit.

Open Science Initiative: Membuat repository terbuka untuk dataset anonim dan kode model (dengan persetujuan etik).

Penelitian ini telah membuktikan potensi hybrid ensemble machine learning untuk prediksi diabetes berbasis data SIMRS. Implementasi lebih lanjut membutuhkan pendekatan holistik yang mempertimbangkan aspek teknis, klinis,

dan organisasional untuk memastikan manfaat nyata bagi sistem kesehatan di Indonesia.

DAFTAR PUSTAKA

[1] Wang et al., "Prediction of Type 2 Diabetes Risk and Its Effect Using Machine Learning Algorithms", *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, 2020, https://doi.org/10.3390/ijerph17062012

[2] Abdurrahman et al., "Optimasi Algoritma XGBoost Classifier Menggunakan GridSearch dan Random Search untuk Prediksi Diabetes", *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 3, 2022, pp. 501-510, https://jtiik.ub.ac.id/index.php/jtiik/article/view/6879

[3] Bandil & Dandotiya, "Hyperparameter Optimization Techniques for Enhanced Diabetes Prediction Using XGBoost", *International Journal of Recent Innovations Trends in Computing and Communication (IJRITCC)*, vol. 11, no. 5, 2023, http://www.ijritcc.org/abstract.php?id=9340

[4] Li W, Peng Y, et al., "Diabetes Prediction Model Based on GA-XGBoost and Stacking Ensemble Algorithm", *PLOS ONE*, vol. 19, no. 4, 2024, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0304716

[5] Zhou et al., "Optimalisasi Model Klasifikasi Diabetes Menggunakan Boruta Feature Selection dan Ensemble Stacking", *Journal of Medical Systems*, vol. 47, no. 3, 2023, pp. 1-15.

[6] Gregorius Airlangga, "Enhancing Diabetes Prediction Accuracy through Hybrid Machine Learning Models: A Comparative Study", *G-Tech : Jurnal Teknologi Terapan*, vol. 8, no. 2, 2024, pp. 1297-1306, E-ISSN: 2623-064X, P-ISSN: 2580-8737

[7] Wibisono et al., "Optimalisasi Model Klasifikasi Diabetes Menggunakan

AdaBoost, Gradient Boosting, dan XGBoost", *JTSiskom*, vol. 12, no. 1, 2024, pp. 45-56, https://jtsiskom.ub.ac.id/index.php/jtsiskom/article/view/543

[8] Kementerian Kesehatan Republik Indonesia, *Laporan Nasional Riskesdas 2018*, Jakarta: Badan Penelitian dan Pengembangan Kesehatan, 2019.

[9] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed., Brussels: International Diabetes Federation, 2021.

[10] Kementerian Kesehatan Republik Indonesia, *Pedoman Nasional Pelayanan Kedokteran Diabetes Melitus Tipe 2*, Jakarta: Kementerian Kesehatan RI, 2022.

LAMPIRAN

Lampiran 1 Dataset

Lampiran 2 Tabel Hasil A