IMPLEMENTASI ALGORITMA RANDOM FOREST UNTUK PREDIKSI RISIKO DIABETES BERBASIS DATA KLINIS

Enos Christean Oimolala Telaumbanua, Susiana
Ilmu Komputer, Universitas Negeri Medan
Jl. William Iskandar Ps. V, Kenangan Baru, Kec. Percut Sei Tuan,
Kabupaten Deli Serdang, Sumatera Utara 20221
enoztel@gmail.com

ABSTRAK

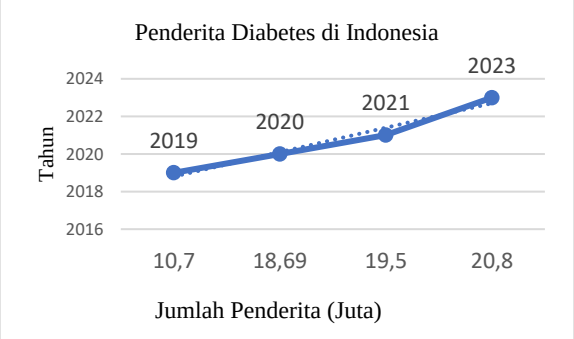
Diabetes merupakan penyakit metabolik kronis yang prevalensinya terus meningkat dan berpotensi menimbulkan komplikasi serius apabila tidak dideteksi sejak dini. Permasalahan utama dalam penanganan diabetes adalah keterbatasan sistem prediksi yang akurat dan mudah digunakan untuk mengidentifikasi risiko pasien berdasarkan data klinis. Penelitian ini bertujuan membangun model prediksi risiko diabetes menggunakan algoritma Random Forest berbasis data pasien RSUD dr. M. Thomsen Nias. Data yang digunakan berjumlah 140 sampel dengan variabel usia, tekanan darah, glukosa darah, kolesterol, HbA1c, pola makan, dan jenis kelamin. Metode penelitian meliputi praproses data, pembangunan pipeline Random Forest, optimasi hyperparameter menggunakan RandomizedSearchCV, serta evaluasi model dengan 5-Fold Cross Validation. Hasil penelitian menunjukkan bahwa model memiliki kinerja yang sangat baik dengan akurasi 93,6%, presisi 94,2%, recall 99,2%, dan F1-score 96,7%. Variabel yang paling berpengaruh adalah glukosa darah sewaktu, glukosa darah puasa, HbA1c, dan usia. Model kemudian diimplementasikan dalam aplikasi berbasis web menggunakan Streamlit untuk memudahkan deteksi dini. Hasil ini menunjukkan bahwa Random Forest efektif dan andal dalam memprediksi risiko diabetes berbasis data klinis.

Kata kunci : *Data Klinis, Machine Learning , Prediksi Diabetes, Random Forest, Streamlit*

1. PENDAHULUAN

Diabetes melitus merupakan penyakit kronis yang ditandai dengan hiperglikemia akibat gangguan produksi atau fungsi insulin. Penyakit ini sering tidak terdeteksi pada tahap awal karena minimnya gejala, namun dapat menimbulkan komplikasi serius seperti penyakit jantung, stroke, gagal ginjal, dan gangguan penglihatan [1], [2]. Selain itu, faktor genetik dan pola hidup tidak sehat turut berkontribusi terhadap peningkatan risiko diabetes [3], [4].

Secara global, jumlah penderita diabetes terus meningkat dan menjadi masalah kesehatan yang signifikan. Indonesia termasuk dalam negara dengan jumlah penderita diabetes yang tinggi dan diproyeksikan terus mengalami peningkatan hingga tahun 2045 [5]. Kondisi ini menunjukkan pentingnya upaya deteksi dini untuk mengurangi risiko komplikasi dan beban sistem kesehatan [6]. Hal ini juga tercermin dari data nasional, di mana berdasarkan grafik "Penderita Diabetes di Indonesia" terlihat adanya tren peningkatan jumlah penderita yang signifikan dari tahun ke tahun. Berdasarkan data Kementerian Kesehatan, pada tahun 2019 jumlah penderita diabetes di Indonesia tercatat sebesar 10,7 juta jiwa, kemudian mengalami lonjakan tajam pada tahun 2020 menjadi 18,69 juta jiwa, yang menunjukkan peningkatan lebih dari 70% hanya dalam satu tahun. Kenaikan ini berlanjut pada tahun 2021 dengan jumlah penderita mencapai 19,5 juta jiwa, dan terus meningkat hingga tahun 2023 dengan angka tertinggi sebesar 20,8 juta jiwa. Secara keseluruhan, dalam rentang waktu lima tahun dari 2019 hingga 2023, jumlah penderita diabetes di Indonesia hampir dua kali lipat. Data ini dapat dilihat pada gambar 1 Data ini menegaskan bahwa diabetes merupakan masalah kesehatan masyarakat yang semakin serius di Indonesia, sehingga diperlukan perhatian khusus melalui upaya pencegahan, deteksi dini, serta penanganan yang didukung oleh pendekatan berbasis data dan teknologi [7]



Gambar 1. Jumlah Penderita Diabetes di Indonesia

Pemanfaatan teknologi machine learning, khususnya teknik klasifikasi dalam data mining, dapat digunakan untuk membantu deteksi dini diabetes berdasarkan data klinis [8]Salah satu algoritma yang memiliki performa baik adalah Random Forest, yang mampu menghasilkan akurasi tinggi serta melakukan seleksi fitur secara otomatis [9], [10]. Namun, permasalahan ketidakseimbangan data sering mempengaruhi kinerja model, sehingga diperlukan metode seperti Synthetic Minority Oversampling Technique (SMOTE) untuk meningkatkan hasil klasifikasi [11].

Peramalan penjualan merupakan faktor penting dalam penentuan kebijakan dan perencanaan produksi di perusahaan, termasuk PT. Perkebunan Nusantara IV (PTPN IV). Penelitian ini bertujuan menentukan model time series terbaik menggunakan metode proyeksi tren untuk meramalkan penjualan Minyak Kelapa Sawit (MKS) periode Maret–Desember 2015. Dengan pendekatan deskriptif dan analisis menggunakan Microsoft Excel, hasil menunjukkan bahwa metode tren linier lebih unggul dibandingkan tren eksponensial karena memiliki tingkat error (SEE) yang lebih rendah, yaitu sebesar 591,34 [12].

Meskipun berbagai penelitian telah dilakukan, sebagian besar masih terbatas pada jumlah variabel yang digunakan dan belum mengkaji secara optimal pengaruh penanganan data tidak seimbang, khususnya pada populasi di Indonesia [13], [14]. Selain itu, implementasi model prediksi ke dalam sistem berbasis web masih belum banyak dikembangkan sebagai alat bantu deteksi dini.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk membangun model prediksi risiko diabetes menggunakan algoritma Random Forest dengan perbandingan penggunaan SMOTE dan tanpa SMOTE. Model yang dihasilkan kemudian diimplementasikan dalam aplikasi berbasis web agar dapat digunakan secara praktis untuk mendukung deteksi dini diabetes..

2. TINJAUAN PUSTAKA

2.1. Diabetes

Diabetes merupakan penyakit metabolik kronis yang disebabkan oleh gangguan sekresi atau fungsi hormon insulin dan ditandai dengan hiperglikemia [15]. Kondisi ini dapat menimbulkan komplikasi serius seperti kerusakan pada jantung, ginjal, mata, dan pembuluh darah apabila tidak ditangani dengan baik. Selain faktor biologis, kejadian diabetes juga dipengaruhi oleh kondisi lingkungan, sosial ekonomi, dan budaya, seperti rendahnya pengetahuan tentang pola hidup sehat, keterbatasan akses layanan kesehatan, serta kebiasaan konsumsi makanan tinggi gula dan lemak [16]. Di Indonesia, diabetes dengan komplikasi menjadi salah satu penyebab kematian tertinggi, yang menunjukkan dampak besar terhadap aspek kesehatan, sosial, dan ekonomi [16].

Risiko diabetes meningkat seiring bertambahnya usia, terutama pada individu di atas 40 tahun yang lebih rentan mengalami resistensi insulin. Selain itu, berbagai faktor lain seperti jenis kelamin, genetik, kurangnya aktivitas fisik, pola makan tidak sehat, konsumsi alkohol, kebiasaan merokok, tekanan darah, kolesterol, stres, dan pola tidur turut mempengaruhi perkembangan penyakit ini [17]. Faktor-faktor tersebut saling berkaitan dalam memperburuk kondisi metabolisme tubuh, sehingga pencegahan diabetes memerlukan pendekatan yang komprehensif melalui intervensi medis serta perubahan gaya hidup dan peningkatan kesadaran masyarakat.

2.2. Klasifikasi

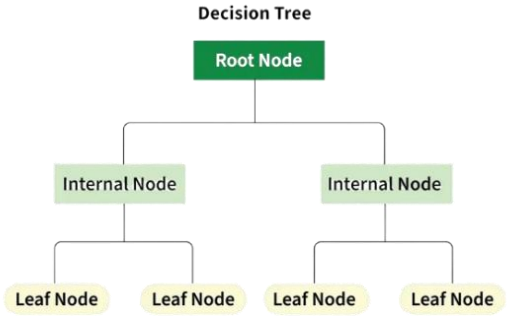
Klasifikasi merupakan metode dalam machine learning yang bertujuan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan fitur yang dimiliki, di mana dalam supervised learning model dibangun menggunakan data berlabel untuk memprediksi kelas pada data baru. Salah satu algoritma yang banyak digunakan adalah Random Forest, yang bekerja dengan menggabungkan sejumlah Decision Tree yang dilatih menggunakan subset acak data (bootstrap sampling) dan fitur, sehingga mampu mengurangi overfitting serta meningkatkan akurasi dan stabilitas prediksi [18]

2.3. Data Mining

Data mining merupakan proses untuk menemukan pola tersembunyi, hubungan, atau informasi bernilai dari data dalam jumlah besar menggunakan metode statistik, matematika, dan algoritma pembelajaran mesin, yang juga dikenal sebagai Knowledge Discovery in Databases (KDD) untuk mendukung pengambilan keputusan [19]. Dalam konteks medis, data mining berperan penting dalam analisis rekam medis, identifikasi faktor risiko, serta pengembangan sistem prediksi diagnosis dini melalui berbagai teknik seperti klasifikasi, klastering, asosiasi, dan deteksi anomali, di mana salah satu metode klasifikasi yang umum digunakan adalah decision tree yang menjadi dasar algoritma Random Forest. Proses data mining meliputi tahapan pemilihan data, pembersihan, transformasi, penerapan algoritma, serta evaluasi hasil, yang sangat penting untuk memastikan kualitas data sehingga dapat menghasilkan informasi yang akurat dan relevan dalam mendukung deteksi dini dan pengambilan keputusan klinis [20]

2.4. Decision Tree

Decision Tree atau pohon keputusan merupakan metode dalam data mining dan machine learning yang digunakan untuk klasifikasi dan prediksi dengan struktur berupa node (kondisi/atribut), cabang (hasil pengujian), dan leaf node (keputusan akhir), sehingga mudah dipahami dan efektif diterapkan, termasuk di bidang kesehatan [21]. Gambar 2 Struktur Decision Tree menunjukkan hubungan antara node, cabang, dan leaf dalam proses pengambilan keputusan. Dalam konteks prediksi risiko diabetes, metode ini mengklasifikasikan individu berdasarkan data klinis seperti usia, BMI, tekanan darah, dan kadar glukosa untuk mengidentifikasi risiko secara sederhana dan mendukung deteksi dini [22]Decision Tree memiliki keunggulan dalam menangani data numerik dan kategorikal serta mudah diinterpretasikan, namun rentan terhadap overfitting sehingga perlu teknik pruning, dan pengembangannya melalui Random Forest dapat meningkatkan akurasi meskipun pemahaman dasar Decision Tree tetap penting [23].



Gambar 2. Struktur *Decision Tree*

2.5. Random Forest

Random Forest merupakan algoritma machine learning berbasis *ensemble learning* yang bertujuan meningkatkan akurasi dan stabilitas prediksi dengan menggabungkan banyak model, baik untuk klasifikasi maupun regresi [24]. Algoritma ini bekerja dengan membangun sejumlah Decision Tree menggunakan teknik *bootstrap sampling*, yaitu pengambilan sampel data secara acak dengan pengembalian, serta pemilihan subset fitur secara acak pada setiap node untuk mengurangi overfitting dan meningkatkan keberagaman model [25]. Setiap pohon dalam *ensemble* memberikan prediksi, kemudian hasil akhir ditentukan melalui *majority voting* untuk klasifikasi [26].

Pengembangan Random Forest berasal dari konsep decision forest yang bertujuan mengatasi kelemahan Decision Tree tunggal, khususnya kesalahan yang dapat menyebar dari node akar hingga daun [27]. Dalam proses klasifikasi, setiap pohon menghasilkan prediksi $h_1(x),h_2(x),...,h_T(x)$, dan hasil akhir ditentukan berdasarkan kelas yang paling banyak dipilih. Sedangkan pada regresi, prediksi diperoleh dari rata-rata output seluruh pohon, sehingga menghasilkan estimasi yang lebih stabil dan akurat, terutama pada data yang mengandung noise atau outlier.

Dalam pembentukan pohon, Random Forest menggunakan fungsi pengukuran impuritas seperti *Gini Impurity* dan *Entropy* untuk menentukan pemisahan terbaik pada setiap node. Pemilihan atribut dilakukan dengan meminimalkan nilai impuritas agar menghasilkan pembagian data yang lebih homogen. Proses ini serupa dengan metode *Classification and Regression Tree* (CART), namun tanpa pruning, serta memanfaatkan pemilihan fitur acak untuk meningkatkan performa model [28]

Secara matematis, *Random Forest* memanfaatkan konsep vektor acak independen (IID) dan fungsi margin untuk mengukur kekuatan klasifikasi serta batas kesalahan generalisasi model [29]. Pendekatan ini memungkinkan kombinasi banyak model lemah (*weak learners*) menjadi model yang lebih kuat (*strong learner*), sehingga mampu mengurangi *overfitting* dan meningkatkan generalisasi.

Keunggulan utama *Random Forest* adalah kemampuannya dalam menangani data kompleks, *robust* terhadap *noise*, serta mampu mengidentifikasi pentingnya fitur (*feature importance*), yang sangat berguna dalam analisis data medis [30]. Penelitian menunjukkan bahwa algoritma ini memiliki performa tinggi dengan akurasi hingga 98,4% dalam klasifikasi data medis, sehingga sangat potensial digunakan dalam sistem deteksi dini penyakit seperti diabetes [31].

Selain itu, *Random Forest* relatif mudah diimplementasikan karena tersedia pada berbagai library seperti Scikit-learn dan Weka, serta tidak terlalu bergantung pada normalisasi data. Namun, pemilihan parameter seperti jumlah pohon dan kedalaman maksimum tetap perlu diperhatikan agar model tidak terlalu kompleks. Dengan keunggulan tersebut, Random Forest menjadi salah satu metode yang efektif dan andal dalam pengembangan sistem prediksi berbasis machine learning di bidang kesehatan [32].

2.6. Confusion Matrix

Confusion matrix merupakan matriks yang digunakan untuk mengevaluasi kinerja model dengan membandingkan kelas aktual dan hasil prediksi, di mana elemen diagonal utama menunjukkan prediksi yang benar dan elemen lainnya menunjukkan kesalahan klasifikasi [33]. Matriks ini terdiri dari *True Positive* (TP) sebagai data positif yang diprediksi benar, *True Negative* (TN) sebagai data negatif yang diklasifikasikan dengan tepat, *False Negative* (FN) sebagai data positif yang salah diprediksi negatif, serta *False Positive* (FP) sebagai data negatif yang salah diprediksi positif [34]. Berdasarkan nilai-nilai tersebut, performa model dapat diukur menggunakan metrik seperti *accuracy, recall, precision*, dan *F1-Score*, di mana *accuracy* menunjukkan tingkat ketepatan keseluruhan, *recall* mengukur kemampuan mendeteksi data positif, *precision* menunjukkan ketepatan prediksi positif, dan *F1-Score* merepresentasikan keseimbangan antara precision dan recall.

2.7. Metrik Evaluasi Kinerja Model

Evaluasi kinerja model bertujuan menilai kemampuan model dalam menghasilkan prediksi yang akurat berdasarkan perbandingan antara hasil prediksi dan data aktual. Dalam penelitian ini digunakan metrik *accuracy, precision, recall,* dan *F1-Score* untuk memberikan gambaran performa model, khususnya pada klasifikasi medis yang sensitif terhadap kesalahan prediksi. *Accuracy* menunjukkan ketepatan keseluruhan, *recall* mengukur kemampuan mendeteksi data positif, *precision* menunjukkan ketepatan prepositif, dan *F1-Score* merupakan keseimbangan antara *precision* dan *recall*, yang perhitungannya ditunjukkan pada Persamaan (1.1) hingga (1.4).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \; x \; 100\% \quad (1.1)$$

$$Recall = \frac{TP}{(TP+FN)} \; x \; 100\% \quad (1.2)$$

$$Precision = \frac{TP}{(TP+FP)} \; x \; 100\% \quad (1.3)$$

$$F1 - Score = \frac{2 \; x \; recall \; x \; precision}{(recall + precision)} \; x \; 100\% \quad (1.4)$$

2.8. Penelitian Terdahulu

Penelitian oleh [4] menunjukkan bahwa algoritma Random Forest memiliki performa tinggi dalam klasifikasi diabetes berdasarkan data rekam medis di RSUP Dr. M. Jamil Padang. Dibandingkan dengan Naive Bayes dan Logistic Regression, model ini mencapai accuracy 96,7%, precision 95,2%, dan recall 94,8%. Hasil tersebut menunjukkan kemampuan model yang akurat dan andal dalam mendeteksi kasus positif diabetes. Selain itu, penelitian ini menekankan pentingnya feature selection, di mana fitur seperti poliuria, polifagia, dan penyembuhan luka lambat memiliki kontribusi besar. Random Forest juga terbukti stabil melalui cross-validation dan tahan terhadap overfitting karena menggunakan pendekatan ensemble, sehingga berpotensi diterapkan dalam sistem pendukung keputusan medis untuk diagnosis dini diabetes.

Penelitian oleh [5] membahas penerapan Random Forest yang dikombinasikan dengan teknik seleksi fitur menggunakan Recursive Feature Elimination (RFE) pada dataset Pima Indian Diabetes. Dengan mengurangi jumlah fitur dari delapan menjadi lima (terutama glukosa, BMI, dan riwayat keluarga), model mampu meningkatkan akurasi dari 93,45% menjadi 97,12%, sekaligus mengurangi kompleksitas dan waktu komputasi. Selain itu, precision dan recall juga mengalami peningkatan, menunjukkan keandalan model dalam klasifikasi. Dibandingkan dengan SVM dan KNN, Random Forest lebih unggul dalam stabilitas dan ketahanan terhadap overfitting. Penelitian ini menyimpulkan bahwa kombinasi Random Forest dan seleksi fitur menghasilkan model yang lebih efisien, akurat, dan mudah diimplementasikan dalam sistem prediksi berbasis web maupun aplikasi diagnosis klinis.

3. METODE PENELITIAN

Penelitian ini dilaksanakan di RSUD dr. M. Thomsen Nias yang berlokasi di Kota Gunungsitoli, Sumatera Utara, selama dua bulan dengan tahapan meliputi pengumpulan data klinis, praproses data, pelatihan model menggunakan algoritma Random Forest, serta implementasi model ke dalam aplikasi web berbasis Streamlit. Jenis penelitian yang digunakan adalah kuantitatif dengan pendekatan eksperimen komputasional berbasis positivisme, di mana data klinis dianalisis menggunakan metode statistik dan machine learning untuk menguji hubungan antar variabel serta mengevaluasi kinerja model prediksi diabetes.

3.1. Populasi dan Sampel

Populasi dalam penelitian ini adalah seluruh pasien yang memiliki riwayat pemeriksaan klinis terkait diabetes melitus di RSUD dr. M. Thomsen Nias, dengan sampel sebanyak 140 data rekam medis yang terdiri dari 132 pasien diabetes melitus (DM) dan 8 pasien non-DM. Klasifikasi dilakukan dengan mengelompokkan nilai 0 sebagai Non-DM, sedangkan DM, DM+HT, dan DM+HT dikategorikan sebagai DM (label = 1) karena menunjukkan kondisi diabetes dengan atau tanpa hipertensi. Teknik sampling yang digunakan adalah simple random sampling, di mana setiap data memiliki peluang yang sama untuk terpilih, kemudian data dibagi secara acak untuk pelatihan dan validasi menggunakan metode 5-Fold Cross Validation guna menghasilkan evaluasi model yang lebih stabil dan representatif

3.2. Teknik Pengumpulan Data

Teknik pengumpulan data dilakukan melalui observasi dan dokumentasi rekam medis pasien, dengan variabel klinis yang dikumpulkan meliputi usia, tekanan darah, pola makan, jenis kelamin, insulin, glukosa darah puasa, HbA1c, kolesterol, glukosa darah sewaktu, indeks massa tubuh (BMI), trigliserida, serta diagnosa sebagai label klasifikasi. Seluruh data pasien telah disamarkan untuk menjaga kerahasiaan dan etika penelitian medis. Data diperoleh dari RSUD dr. M. Thomsen Nias dan disimpan dalam format Microsoft Excel (.xlsx) dengan total 140 data pasien dan 13 fitur klinis, di mana variabel DIAGNOSA digunakan sebagai target yang dikonversi menjadi biner, yaitu 1 untuk pasien diabetes melitus (DM) dan 0 untuk non-DM.

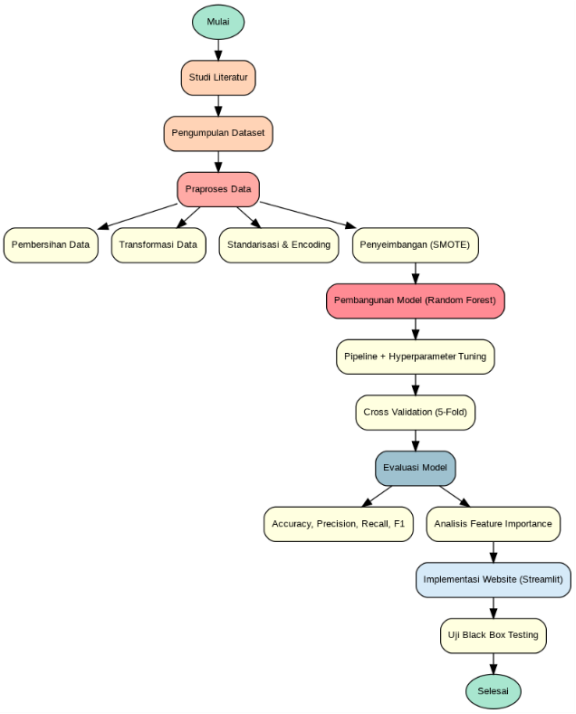
Tabel 1. *Dataset* Klinis

No	Variabel Klinis	Keterangan
1	Usia	Usia pasien dalam tahun
2	Tekanan Darah (mmHg)	Tekanan darah sistolik/diastolik
3	Pola Makan	Frekuensi makan per hari
4	Jenis Kelamin	Laki-laki / Perempuan
5	Insulin	Jenis dan kadar insulin pasien
6	Glukosa Darah Puasa (mg/dL)	Kadar glukosa sebelum makan
7	HbA1c (%)	Persentase kadar hemoglobin terglikasi
8	Kolesterol (mg/dL)	Kadar kolesterol total
9	Glukosa Darah Sewaktu (mg/dL)	Kadar glukosa darah secara acak
10	BB/TB (BMI)	Indeks massa tubuh pasien
11	Trigliserida	Kadar lemak darah
12	Diagnosa	Label hasil diagnosa (DM / Tidak DM)

3.3. Prosedur Penelitian

Prosedur penelitian ini terdiri dari beberapa tahap seperti pada Gambar 1 (Diagram Alir Proses Penelitian), yaitu: (A) studi literatur untuk memahami

teori diabetes, algoritma Random Forest, dan penelitian terkait; (B) pengumpulan dataset dari rekam medis RSUD dr. M. Thomsen Nias sebanyak 140 data dengan 13 fitur klinis, dengan variabel target DIAGNOSA (1 = DM, 0 = non-DM); (C) praproses data meliputi pembersihan data, imputasi nilai kosong, pemisahan tekanan darah menjadi sistolik dan diastolik, encoding data kategorikal, standarisasi, serta penyeimbangan data menggunakan SMOTE; (D) pembangunan model menggunakan Random Forest dengan pipeline, tuning hyperparameter menggunakan RandomizedSearchCV, dan validasi 5-Fold Stratified Cross Validation dengan metrik F1-Score; (E) evaluasi model menggunakan accuracy, precision, recall, F1-score, serta analisis feature importance; dan (F) implementasi model ke dalam website berbasis Streamlit untuk prediksi risiko diabetes, termasuk input data pasien, hasil prediksi, dan pengujian sistem dengan black box testing.

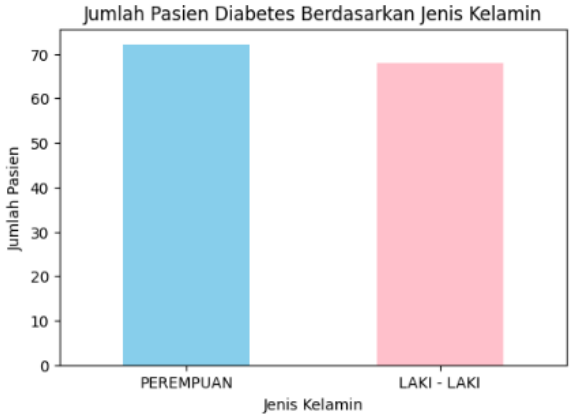


Gambar 3. Diagram Alir Proses Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Data

Dataset penelitian ini berisi data klinis pasien yang mencakup aspek demografis, fisiologis, gaya hidup, dan hasil laboratorium. Data demografis meliputi usia (numerik) dan jenis kelamin (laki-laki/perempuan), yang berpengaruh terhadap risiko diabetes akibat faktor biologis dan metabolisme [35]. Parameter fisiologis mencakup tekanan darah serta berat dan tinggi badan untuk menilai hipertensi dan indeks massa tubuh, sedangkan pola makan menggambarkan kebiasaan konsumsi yang memengaruhi kadar glukosa dan lipid [36]. Parameter laboratorium meliputi glukosa darah puasa, glukosa sewaktu, HbA1c, insulin, kolesterol, dan trigliserida sebagai indikator utama kondisi metabolik. Dari 140 pasien, terdapat 72 perempuan (51,4%) dan 68 laki-laki (48,6%) dengan distribusi yang relatif seimbang. Gambar 4 menunjukkan perbandingan jumlah pasien berdasarkan jenis kelamin, yang mendukung analisis risiko diabetes secara adil



Gambar 4. Jumlah Pasien Diabetes Berdasarkan Jenis Kelamin

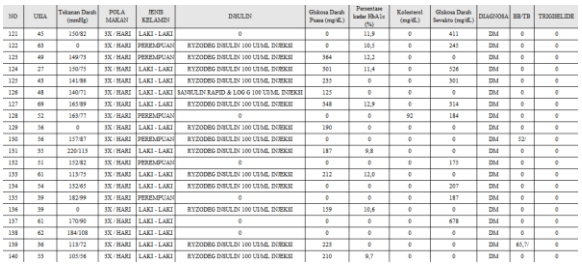
4.2. Pengumpulan *Dataset*

Dataset penelitian diperoleh dari rekam medis pasien di RSUD dr. M. Thomsen Nias secara sekunder, dengan total 140 sampel dan 13 atribut yang mencakup data demografis, fisiologis, gaya hidup, dan parameter laboratorium. Atribut yang digunakan meliputi usia, jenis kelamin, pola makan, insulin, tekanan darah (sistolik dan diastolik), BMI (BB/TB), glukosa darah puasa dan sewaktu, HbA1c, kolesterol, serta trigliserida. Indikator klinis seperti glukosa, HbA1c, dan tekanan darah digunakan karena mampu menggambarkan kondisi metabolik jangka pendek dan panjang, sedangkan kolesterol dan trigliserida berkaitan dengan risiko kardiovaskular. Pemecahan tekanan darah menjadi BP_SYSTOLIC dan BP_DIASTOLIC dilakukan untuk meningkatkan ketelitian analisis. Kombinasi atribut ini diharapkan memberikan representasi kondisi pasien secara komprehensif sehingga model Random Forest dapat mengenali pola dan memprediksi risiko diabetes dengan akurasi tinggi [37]. Tabel 2 menjelaskan atribut yang digunakan, sedangkan Gambar 5 menampilkan struktur dataset yang terdiri dari data numerik dan kategorikal.

Tabel 2. Atribut *Dataset* Penelitian

No	Nama Atribut	Keterangan
1	Usia	Umur pasien (tahun)
2	Tekanan Darah (mmHg)	Nilai tekanan darah sistolik/diastolik pasien
3	Pola Makan	Frekuensi makan pasien per hari
4	Jenis Kelamin	Laki-laki atau perempuan
5	Insulin	Jenis atau keberadaan penggunaan insulin

No	Nama Atribut	Keterangan
6	Glukosa Darah Puasa (mg/dL)	Kadar glukosa dalam darah setelah puasa
7	Persentase Kadar HbA1c (%)	Rata-rata kadar gula darah dalam 3 bulan terakhir
8	Kolesterol (mg/dL)	Kadar kolesterol total dalam darah
9	Glukosa Darah Sewaktu (mg/dL)	Kadar glukosa darah pada waktu acak
10	BB/TB	Indeks massa tubuh (BMI) pasien
11	Trigiselide	Kadar trigliserida dalam darah
12	BP_SYSTOLIC	Tekanan darah sistolik hasil pemecahan atribut
13	BP_DIASTOLIC	Tekanan darah diastolik hasil pemecahan atribut



NO	USIA	Tekanan Darah	JENIS KELAMIN	INSULIN	Glukosa Darah Puasa (mg/dL)	Persentase kadar HbA1c (%)	Kolesterol (mg/dL)	Glukosa Darah Sewaktu (mg/dL)	DIAGNOSA	BB/TB	TRIGISELIDE
121	45	150/82	3X / HARI	LAKI - LAKI	0	0	11,9	0	411	DM	0	0
122	63	0	3X / HARI	PEREMPUAN	RYZODEG INSULIN 100 U/ML INJEKSI	0	10,3	0	245	DM	0	0
123	49	149/73	3X / HARI	PEREMPUAN	RYZODEG INSULIN 100 U/ML INJEKSI	364	12,2	0	0	DM	0	0
124	27	150/75	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	301	11,4	0	526	DM	0	0
125	45	141/86	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	255	0	0	301	DM	0	0
126	48	140/71	3X / HARI	LAKI - LAKI	SANSULIN RAPID & LOG G 100 U/ML INJEKSI	125	0	0	0	DM	0	0
127	69	165/89	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	348	12,9	0	314	DM	0	0
128	52	163/77	3X / HARI	PEREMPUAN	0	0	0	92	184	DM	0	0
129	56	0	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	190	0	0	0	DM	0	0
130	56	157/87	3X / HARI	PEREMPUAN	RYZODEG INSULIN 100 U/ML INJEKSI	0	0	0	0	DM	32	0
131	35	220/113	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	187	9,8	0	0	DM	0	0
132	51	152/82	3X / HARI	PEREMPUAN	0	0	0	0	175	DM	0	0
133	61	113/75	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	212	12,0	0	0	DM	0	0
134	54	152/65	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	0	0	0	207	DM	0	0
135	39	182/99	3X / HARI	PEREMPUAN	0	0	0	0	187	DM	0	0
136	59	0	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	159	10,6	0	0	DM	0	0
137	61	170/90	3X / HARI	LAKI - LAKI	0	0	0	0	678	DM	0	0
138	62	184/108	3X / HARI	LAKI - LAKI	0	0	0	0	0	DM	0	0
139	36	113/72	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	223	0	0	0	DM	60,7	0
140	53	103/56	3X / HARI	LAKI - LAKI	RYZODEG INSULIN 100 U/ML INJEKSI	210	9,7	0	0	DM	0	0

Gambar 5. Tampilan Data Klinis Untuk *Dataset*

4.3. Penyeimbangan Data : Teknik SMOTE

Dataset awal tidak seimbang (DM = 132, Non-DM = 8) sehingga berpotensi menyebabkan bias pada model. Untuk mengatasinya digunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*) yang menghasilkan data sintetis pada kelas minoritas melalui interpolasi, bukan duplikasi, sehingga variasi data tetap terjaga [38]. Proses SMOTE dilakukan dengan memilih data minoritas dan tetangga terdekatnya, kemudian membentuk data baru menggunakan rumus:

$$x_{new} = x_i + \delta(x_{nn} - x_i) \qquad (1.5)$$

Dengan x_i = data minoritas, x_{nn} = tetangga terdekat, dan $\delta \in [0,1]$. Contoh: usia 44 dan 50 menghasilkan data baru usia 47, serta glukosa 187 dan 180 menghasilkan 183,5 mg/dL. Setelah SMOTE, data menjadi seimbang (DM = 132, Non-DM = 132). Penerapan SMOTE dilakukan dalam pipeline bersama Random Forest dan validasi cross-validation sehingga hanya diterapkan pada data training, mencegah data leakage dan menjaga keakuratan evaluasi [39].

4.4. Praproses Data

Praproses data dilakukan untuk memastikan dataset siap digunakan oleh model dengan mengatasi missing values, ketidaksesuaian data, dan ketidakseimbangan kelas [40]. Pada tahap data cleaning, dilakukan penyeragaman label diagnosis seperti "DM", "DM+HT", dan lainnya menjadi biner (1 = diabetes, 0 = non-diabetes), serta penghapusan kolom yang tidak relevan seperti kolom NO karena hanya sebagai identitas. Data yang tidak valid atau tidak teridentifikasi juga dihapus untuk menghindari bias pada model [41]. Gambar 6 menunjukkan proses penghapusan kolom NO.

NO	USIA	Tekanan Darah (mmHg)	POLA MAKAN
121	45	150/82	3X / HARI
122	63	0	3X / HARI
123	49	149/73	3X / HARI
124	27	150/75	3X / HARI

Gambar 6. Penghapusan Kolom NO

Imputasi data dilakukan untuk menangani nilai kosong dan nol pada atribut klinis seperti glukosa, HbA1c, kolesterol, BB/TB, dan trigliserida. Pada atribut numerik, nilai yang hilang digantikan dengan median untuk mengurangi pengaruh nilai ekstrem [42], dengan rumus:

$$Median = \frac{x_{\frac{n}{2}} + x_{(\frac{n}{2})+1}}{2}$$

Contoh, atribut usia memiliki median 54 tahun sehingga nilai kosong diganti dengan 54. Sementara itu, pada atribut kategorikal seperti pola makan, jenis kelamin, dan insulin, nilai kosong diganti dengan label "*MISSING*" agar informasi tetap terjaga dalam proses pemodelan.

Kolom tekanan darah yang awalnya berbentuk string "*sistolik/diastolik*" (misalnya 120/80) diubah menjadi dua kolom numerik agar dapat diproses oleh model. Proses ini meliputi deteksi format "a/b", pemisahan nilai menjadi *BP_SYSTOLIC* (sebelum /) dan *BP_DIASTOLIC* (setelah /), konversi ke tipe numerik (float/int), serta penghapusan kolom asli. Contohnya, 120/80 diubah menjadi BP_SYSTOLIC = 120 dan BP_DIASTOLIC = 80. Transformasi ini penting agar algoritma *machine learning* dapat mengolah data secara optimal.

Target variabel *DIAGNOSA* dikonversi menjadi bentuk biner (1 = DM, 0 = Non-DM) melalui proses pembersihan data, dengan distribusi awal menunjukkan jumlah pasien DM lebih dominan [42]. Gambar 7 menunjukkan proses konversi biner tersebut. Selanjutnya, data kategorikal diubah menggunakan *One-Hot Encoding* karena tidak dapat langsung diproses oleh Random Forest. Jika suatu atribut memiliki kkk kategori, maka akan diubah menjadi kkk variabel biner. Contohnya, atribut jenis kelamin {Laki-laki, Perempuan} direpresentasikan menjadi [1,0] dan [0,1] seperti pada Tabel 3

```
def map_diagnosa(value):
    v = str(value).upper().strip()
    if "DM" in v or "DIABET" in v:
        return 1
    elif "TIDAK" in v and "DM" in v:
        return 0
    elif v in ["-", "NAN", "NONE", ""]:
        return np.nan
    else:
        return 0
```

→

```
Distribusi label (diagnosa_bin):
diagnosa_bin
1    132
0      8
Name: count, dtype: int64
```

Gambar 7. Konversi Bentuk Biner

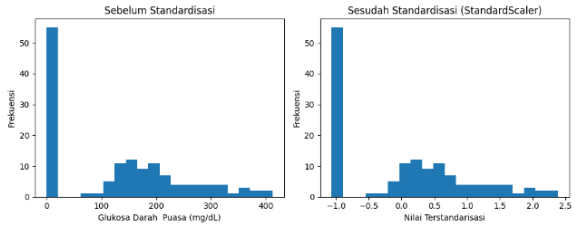
Tabel 3. *Encoding* Data Kategorikal

Laki-laki	Perempuan
1	0
0	1

Standardisasi fitur dilakukan untuk menyamakan skala data numerik seperti usia, glukosa, HbA1c, kolesterol, trigliserida, serta tekanan darah agar memiliki kontribusi yang seimbang dalam model. Metode *StandardScaler* mengubah data sehingga memiliki mean 0 dan standar deviasi 1, menggunakan rumus:

$$X' = \frac{X - \underline{X}}{\sigma}$$

dengan x = nilai asli, μ = rata-rata, σ = standar deviasi dan z = nila hasil standardisasi [43]. Proses ini penting karena perbedaan skala antar fitur dapat menyebabkan bias pada model. Contohnya, nilai glukosa 312 mg/dL menghasilkan nilai standardisasi 1,29, yang berarti berada 1,29 standar deviasi di atas rata-rata. Standardisasi diterapkan melalui pipeline hanya pada fitur numerik untuk mencegah data leakage, sedangkan fitur kategorikal diproses dengan One-Hot Encoding. Gambar 7 menunjukkan perubahan distribusi data sebelum dan sesudah standardisasi



Gambar 7. Perbandingan Distribusi Fitur Glukosa Darah Puasa Sebelum dan Sesudah Standardisasi

4.5. Pembangunan Model Random Forest

Pada penelitian ini digunakan algoritma Random Forest untuk klasifikasi risiko Diabetes Melitus karena kemampuannya menangani banyak fitur dan mengurangi overfitting melalui metode ensemble, dengan alur proses ditunjukkan pada Gambar 8 dan arsitektur model pada Gambar 9 Dataset terdiri dari 140 pasien dengan 13 atribut klinis yang direpresentasikan sebagai matriks fitur. roses dimulai dengan bootstrap sampling, di mana probabilitas data tidak terambil adalah:

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0{,}368$$

Sehingga sekitar 63,2% (≈88 data) digunakan sebagai data latih dan 36,8% (≈52 data) sebagai out-of-bag. Setiap subset digunakan untuk membangun decision tree dengan pemisahan menggunakan Gini Index

$$Gini = 1 - ({p_1}^2 + {p_2}^2)$$

Pada root node diperoleh :

$$p(DM) = \frac{132}{140} = 0{,}9429, \qquad p(Non - DM) = \frac{8}{140} = 0{,}0571$$

$$Gini = 1 - (0{,}9429^2 + 0{,}0571^2) = 0{,}1078$$

Seteleha SMOTE :

$$Gini = 1 - (0{,}5^2 + 0{,}5^2) = 0{,}5$$

Evaluasi model menggunakan 5-Fold Cross Validation dengan contoh perhitungan pada satu fold

$$Accuracy = (27 + 0)/(27 + 0 + 1 + 0) = 0{,}9643$$

$$Precision = \frac{27}{(27 + 1)} = 0{,}9643$$

$$Recall = 27/(27 + 0) = 1{,}0000$$

$$F1 = 2 \times \frac{0{,}9643 \times 1{,}0000}{0{,}9643 + 1{,}0000} \approx 0{,}9818$$

Analisis feature importance dihitung dari penurunan impurity;

$$\Delta Gini = 0{,}4709 - 0{,}3398 = 0{,}1311$$

$$Kontribusi = (94/264) \times 0{,}1311 = 0{,}0467$$

Nilai akhir fitur Glukosa Darah Puasa diperoleh sebesar 0,1538 setelah dirata-ratakan dari 300 pohon. Prediksi akhir dilakukan dengan voting mayoritas atau rata-rata probabilitas :

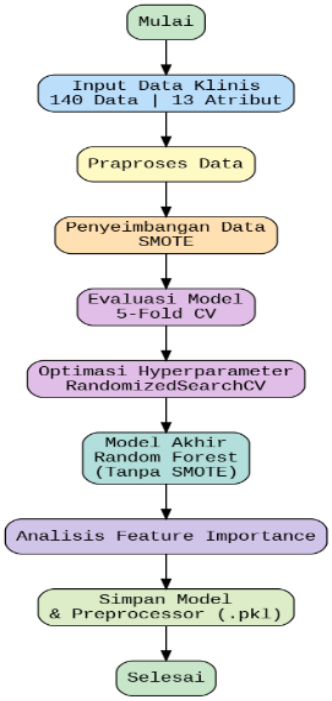
$$\hat{y} = mode\{T_1(x), \ldots, T_k(x)\}$$

$$\hat{y} = 0{,}81 + 0{,}74 + 0{,}61 + 0{,}68 + 0{,}65/5 = 0{,}698$$

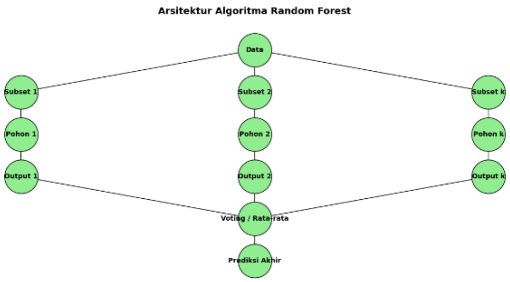
Hasil optimasi hyperparameter menggunakan RandomizedSearchCV menghasilkan parameter terbaik (n_estimators = 300, max_depth = 8, min_samples_split = 8, min_samples_leaf = 6, max_features = 0,2) dengan F1-score ≈ 0,967. Model akhir membentuk fungsi prediksi:

$$\hat{f}(x) = \frac{1}{300} \sum \quad h_i(x)$$

Dengan contoh hasil prediksi sebesar 0,826 (82,6%) yang diklasifikasikan sebagai Diabetes, sehingga menunjukkan bahwa model memiliki performa yang akurat, stabil, dan siap diimplementasikan.



Gambar 8. *Flowchart* Pembangunan Model *Random Forest*



Gambar 9. Arsitektur Algoritma Random Forest

4.6. Analilsi Hasil

Analisis hasil bertujuan untuk menafsirkan kinerja model Random Forest serta mengidentifikasi faktor klinis yang paling berpengaruh melalui analisis feature importance dan evaluasi performa model. *Feature importance* dihitung berdasarkan penurunan impurity menggunakan rumus:

$$FI(X_j) = \frac{1}{T}\sum_{t=1}^{T} \sum_{n \in N_t(j)} \frac{N_n}{N} \Delta i(n)$$

Dengan contoh perhitungan pada fitur Glukosa Darah Puasa:

$$\Delta i(n) = 0{,}4709 - 0{,}3398 = 0{,}1311$$

$$Kontribusi = (\frac{94}{264}) \times 0{,}1311 = 0{,}0467$$

Nilai akhir setelah dirata-ratakan dari $T = 300$ pohon diperoleh:

$$FI = 0{,}1538$$

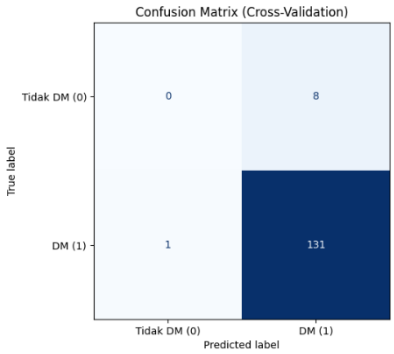
Berdasarkan Tabel 4, fitur paling dominan adalah Glukosa Darah Sewaktu (0,194), diikuti Glukosa Darah Puasa (0,157), Usia (0,122), dan Tekanan Darah Sistolik (0,121), yang menunjukkan bahwa parameter laboratorium memiliki pengaruh utama dalam klasifikasi. Evaluasi model menggunakan 5-Fold Cross Validation (Tabel 5) menghasilkan performa tinggi dengan rata-rata accuracy 0,925, precision 0,944, recall 0,978, dan F1-score 0,959. Berdasarkan confusion matrix (Gambar 10 dan Tabel 1.8), diperoleh perhitungan:

Tabel 4. Urutan 10 Fitur Terpenting Berdasarkan Model *Random Forest*

No	Fitur Klinis	Nilai *Importance*
1	Glukosa Darah Sewaktu (mg/dL)	0.194
2	Glukosa Darah Puasa (mg/dL)	0.157
3	Usia (Tahun)	0.122
4	Tekanan Darah Sistolik (*BP_SYSTOLIC*)	0.121
5	Persentase kadar HbA1c (%)	0.111
6	Tekanan Darah Diastolik (*BP_DIASTOLIC*)	0.109
7	Jenis Insulin (RYZODEG 100 UI/ML)	0.042
8	Jenis Kelamin (Perempuan)	0.040
9	Jenis Kelamin (Laki-laki)	0.032
10	Jenis Insulin (Non-Aktif)	0.024

Tabel 5. Hasil Evaluasi Model *Random Forest* (5-*Fold Cross Validation*)

No	Metrik	Rata-rata	Standar Deviasi
1	Akurasi	0.925	0.026
2	Presisi	0.944	0.035
3	*Recall*	0.978	0.044
4	*F1-Score*	0.959	0.014



Gambar 10. *Confusion Matrix* Model *Random Forest*

Tabel 6. Metrik Evaluasi

No		Prediksi Negatif (0)	Prediksi Positif (1)
1	Aktual Non-DM (0)	0	8
2	Aktual DM (1)	1	131

$$Accuracy = (131 + 0)/(132 + 0 + 8 + 1) = 0{,}9643$$

$$Precision = \frac{131}{(131 + 8)} = 0{,}942$$

$$Recall = 131/(131 + 1) = 0{,}992$$

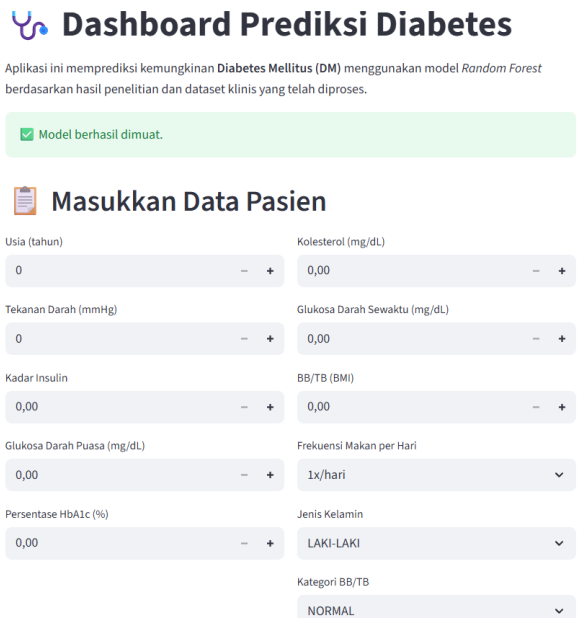
$$F1 = 2 \times \frac{0{,}942 \times 0{,}992}{0{,}942 + 0.992} \approx 0{,}967$$

Hasil ini menunjukkan bahwa model mampu mengenali hampir seluruh pasien diabetes dengan tingkat kesalahan yang sangat rendah, di mana nilai recall yang tinggi menandakan tidak adanya kesalahan dalam mendeteksi pasien DM (*false negative*), sementara precision yang tinggi menunjukkan jumlah false positive yang kecil. Secara keseluruhan, model *Random Forest* memiliki performa yang sangat baik, stabil, dan memiliki kemampuan generalisasi yang tinggi, sehingga layak digunakan sebagai alat bantu prediksi risiko diabetes berbasis data klinis

4.7. Implementasi Website Prediksi Diabetes

Implementasi dilakukan dengan membangun aplikasi web berbasis Streamlit yang mengintegrasikan model Random Forest terbaik hasil pelatihan, tuning, dan cross-validation dalam bentuk file *.pkl* dengan nama *rf_tuned_model_preproc.pkl* untuk memprediksi risiko Diabetes Melitus secara interaktif. Aplikasi ini memungkinkan pengguna memasukkan data klinis pasien melalui formulir input, seperti usia, tekanan darah, kadar glukosa, HbA1c, kolesterol, hingga jenis kelamin, yang kemudian diproses oleh preprocessor sebelum dilakukan prediksi. Sistem juga menyediakan pengaturan

ambang batas (threshold) untuk menentukan sensitivitas hasil, serta menampilkan output berupa probabilitas dan kategori hasil (positif atau negatif DM). Selain itu, aplikasi dilengkapi dengan tampilan awal (Gambar 1.8), tombol prediksi, serta fitur error handling untuk memastikan kemudahan penggunaan. Secara keseluruhan, implementasi ini mempermudah pengguna non-teknis dalam mengakses dan memanfaatkan model prediksi diabetes secara cepat melalui browser tanpa perlu menjalankan kode secara manual.



Gambar 11. Tampilan Awal *Website*

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian "Implementasi Algoritma Random Forest untuk Prediksi Risiko Diabetes Berbasis Data Klinis", dapat disimpulkan bahwa model Random Forest memiliki performa yang sangat baik dan andal dalam memprediksi risiko diabetes, dengan nilai rata-rata F1-Score sebesar 0,967, akurasi 93,6%, presisi 94,2%, dan recall 99,2% pada 5-Fold Cross Validation, yang menunjukkan kemampuan generalisasi tinggi dan sangat efektif dalam mendeteksi pasien DM. Faktor klinis yang paling berpengaruh dalam prediksi meliputi glukosa darah sewaktu, glukosa darah puasa, usia, tekanan darah sistolik, dan HbA1c, yang menegaskan pentingnya parameter metabolik dalam diagnosis diabetes. Implementasi model ke dalam aplikasi web berbasis Streamlit juga berhasil dengan baik dan menghasilkan sistem yang interaktif serta fungsional sebagai alat bantu deteksi dini. Adapun saran untuk pengembangan selanjutnya adalah memperluas dan memperkaya dataset agar lebih representatif, melakukan perbandingan dengan algoritma lain seperti XGBoost, SVM, atau Neural Network, meningkatkan transparansi model menggunakan metode Explainable AI seperti SHAP atau LIME, serta mengembangkan fitur aplikasi agar dapat terintegrasi dengan sistem informasi rumah sakit dan digunakan secara lebih luas dalam praktik klinis

DAFTAR PUSTAKA

[1] W. Apriliah, I. Kurniawan, M. Baydhowi, And T. Haryati, "Prediksi Kemungkinan Diabetes Pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Sistemasi*, Vol. 10, No. 1, P. 163, 2021, Doi: 10.32520/Stmsi.V10i1.1129.

[2] Nurlan, "Perbandingan Pemeberian Eksrak Kulit Manggis Dengan Glibenklamid Terhadap Penurunan Kadar Glukosa Darah Pada Mencit," Vol. 3, No. 2, Pp. 123–129, 2023.

[3] A. N. Faida, Y. Dyah, And P. Santik, "Higeia Journal Of Public Health," Vol. 4, No. 1, Pp. 33–42, 2020.

[4] F. Yusnanda, R. K. Rochadi, And L. T. Maas, "Pengaruh Riwayat Keturunan Terhadap Kejadian Diabetes Mellitus Pada Pra Lansia Di Blud Rsud Meuraxa Kota Banda Aceh Tahun 2017 The Effect Of Heritage History On Dijet Events Of Diabet Mellitus In Pre-Scars In Blud Rsud Meuraxa Kota Banda Aceh 2017," Vol. 4, No. 1, Pp. 18–28, 2018.

[5] E. Zaininda And D. Utama, "Hubungan Antara Kadar Hba1c Dan Kadar Serum Kreatinin Dengan Kejadian Hipertensi Pada Pasien Diabetes Melitus Tipe 2 Di Rumah Sakit Umum Darmayu Ponorogo," Vol. 15, No. 2, Pp. 1–11, 2023.

[6] A. Setiawan, Z. H. Nst, Z. Khairi, And L. Efrizoni, "Klasifikasi Tingkat Risiko Diabetes Menggunakan Algoritma," Vol. 7, No. 2, Pp. 263–271, 2024.

[7] Kemenkes, "Analisis Proksimat, Kandungan Gula Total Dan Uji Organoleptik Beras Analog Gaogu Sebagai Alternatif Pangan Ramah Diabetes," 2019.

[8] H. M. Nawawi, A. B. Hikmah, A. Mustopa, And G. Wijaya, "Model Klasifikasi Machine Learning Untuk Prediksi Ketepatan Penempatan Karir," Vol. 14, No. 1, Pp. 13–25, 2024.

[9] E. P. Febtiawan, L. A. Syamsul, I. Akbar, And A. S. Rachman, "Forecasting Produksi Energi Photovoltaic Menggunakan Algoritma Random Forest Classification," Vol. 5, No. 4, Pp. 1053–1062, 2024, Doi: 10.47065/Josh.V5i4.5514.

[10] A. Fauzi, N. Maulidah, R. Supriyadi, And H. Nalatissifa, "Journal Of Artificial Intelligence And Digital Business (Riggs) Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random," *Journal Of Artificial Intelligence And Digital Business (Riggs)*, Vol. 4, No. 1, Pp. 43–49, 2025.

[11] D. A. Hadi And D. A. N. Sirodj, "Metode Random Forest Untuk Klasifikasi Penyakit Diabetes," In *Bandung Conference Series: Statistics*, 2023, Pp. 428–435.

[12] Susiana, "Analisis Peramalan Penjualan Minyak Kelapa Sawit (Mks) Pada Pt.Perkebunan Nusantara Iv Unit Kebun Pabatu," *Karismatika*, Vol. 1, No. 1, Pp. 43–53, 2015.

[13] R. Maulana, M. F. Hasan, F. Raehan, And M. Ramzy, "Literature Review : Penerapan Algoritma Random Forest Untuk Klasifikasi Penyakit Diabetes," Vol. 2, No. 3, Pp. 550–555, 2024.

[14] C. Z. V. Junus, T. Tarno, And P. Kartikasari, "Klasifikasi Menggunakan Metode Support Vector Machine Dan Random Forest Untuk Deteksi Awal Risiko Diabetes Melitus," *Jurnal Gaussian*, Vol. 11, No. 3, Pp. 386–396, 2023, Doi: 10.14710/J.Gauss.11.3.386-396.

[15] V. D. Ariwati *Et Al.*, "Pendidikan Kesehatan Tentang Diabetes Melitus Pada Masyarakat Rt 3 Kelurahan Curug, Kota Depok," *Jurnal Abdimas-Hip Pengabdian Kepada Masyarakat*, Vol. 4, No. 1, Pp. 47–54, 2023, Doi: 10.37402/Abdimaship.Vol4.Iss1.217.

[16] A. A. Yustian, A. Rahman, A. Fitria, And A. Y. Hariyanto, "Pemberian Informasi Tentang Diabetes Melitus," Vol. 1, Pp. 136–140, 2023.

[17] R. A. S. Kabosu, A. A. Adu, And I. A. T. Hinga, "Faktor Risiko Kejadian Diabetes Melitus Tipe Dua Di Rs Bhayangkara Kota Kupang," *Timorese Journal Of Public Health*, Vol. 1, No. 1, Pp. 11–20, 2019, Doi: 10.35508/Tjph.V1i1.2122.

[18] F. Sodik, B. Dwi, And I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning Pada Data Bank Customers Menggunakan Python," Vol. 3, Pp. 689–694, 2020.

[19] D. Juliadi, B. Irawan, A. Bahtiar, And O. Nurdiawan, "Penerapan Algoritma Fp-Growt Dan Association Rules Pada Pola Pembelian Pizza Hut," Vol. 7, No. 6, Pp. 3443–3448, 2023.

[20] F. S. Wijaya, M. F. Arotsid, And R. Adriano, "Literatur Review : Penerapan Algoritma Random Forest Untuk Klasifikasi Penyakit Diabetes," Vol. 2, No. 3, Pp. 474–481, 2024.

[21] D. Sartika, D. I. Sensuse, U. Indo, G. Mandiri, And F. I. Komputer, "Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , Dan Decision Tree Pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian," Vol. 1, No. 2, Pp. 151–161, 2017.

[22] A. R. Dana, R. V. Kristananda, M. Bagas, S. Wibowo, And D. A. Prasetya, "Perbandingan Algoritma Decision Tree Dan Random Forest Dengan Hyperparameter Tuning Dalam Mendeteksi Penyakit Stroke," Vol. 4, Pp. 66–75, 2024.

[23] R. Fauzan, A. V. Vitianingsih, And D. Cahyono, "Application Of Classification Algorithms In Machine Learning For Phishing Detection," Vol. 5, No. April, Pp. 531–540, 2025.

[24] A. Maehendrayuga And A. Setyanto, "Analisa Prediksi Turnover Karyawan Menggunakan Machine Learning," Vol. 7, No. 2, 2024, Doi: 10.32877/Bt.V7i2.1999.

[25] F. Elfaladonna And A. Rahmadani, "Analisa Metode Classification-Decission Tree Dan Algoritma," Vol. 2, No. 1, Pp. 10–17, 2019.

[26] S. Amaliah, M. Nusrang, And Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng," Vol. 4, No. 2, Pp. 121–127, 2022, Doi: 10.35580/Variansiunm31.

[27] D. Hartama And N. Amalya, "Perbandingan Algoritma Decision Tree , Id3 , Dan Random Forest Dalam Klasifikasi Faktor-Faktor Yang Mempengaruhi," Vol. 6, No. 1, Pp. 72–80, 2025.

[28] D. Alita And A. Rahman, "Pendeteksian Sarkasme Pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," Vol. 8, No. 2, Pp. 50–58, 2020.

[29] L. E. O. Breiman, "Random Forests," Pp. 5–32, 2001.

[30] R. Supriyadi, W. Gata, N. Maulidah, And A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, Vol. 13, No. 2, Pp. 67–75, 2020, Doi: 10.51903/E-Bisnis.V13i2.247.

[31] L. Sari, A. Romadloni, And R. Listyaningrum, "Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, Vol. 14, No. 1, Pp. 155–162, 2023, Doi: 10.35970/Infotekmesin.V14i1.1751.

[32] D. P. Sinambela, H. Naparin, M. Zulfadhilah, And N. Hidayah, "Implementasi Algoritma Decision Tree Dan Random Forest Dalam Prediksi Perdarahan Pascasalin," Vol. 5, No. 3, Pp. 58–64, 2023, Doi: 10.60083/Jidt.V5i3.393.

[33] I. Hakim *Et Al.*, "Klasifikasi Kualitas Produk Mesin Pertanian Berdasarkan Evaluasi Kinerja Algoritma Random Forest," Vol. 6, No. 1, Pp. 343–356, 2025.

[34] R. Nurhidayat And K. E. Dewi, "Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek," Vol. 12, No. 1, Pp. 91–100, 2023.

[35] N. Susanti, P. M. Syifa Rizqi, S. Dewi, And W. Barokah, "Hubungan Jenis Kelamin Terhadap Diabetes Melitus Di Desa Air Hitam," *Jurnal Kesehatan Tambusai*, Vol. 5, No. September, Pp. 7484–7491, 2024.

[36] Muhajiriansyah And R. S. D. Binuko, "Hubungan Antara Kadar Hba1c Dan Gula Darah Sewaktu (Gds) Dengan Kejadian Hipertensi Pada Pasien Diabetes Melitus Tipe Ii Di Rumah Sakit Darmayu Ponorogo," *Health Information : Jurnal Penelitian*, Vol. 15, No. 2, Pp. 1–7, 2023.

[37] M. R. Zuhri, Kusrini, And D. Ariatmanto, "Analisis Perbandingan Algoritma Klasifikasi Untuk Identifikasi Diabetes Dengan Menggunakan Metode Random Forest Dan Naïve Bayes," Pp. 11–20, 2025.

[38] Erlin, Y. Desnelita, N. Nasuton, L. Suryati, And F. Zoromi, "Dampak Smote Terhadap Kinerja Random Forest Classifier Berdasarkan Data Tidak Seimbang," 2022.

[39] C. Haryawan And Y. M. K. Ardhana, "Analisa Perbandingan Teknik Oversampling Smote," *Jire (Jurnal Informatika & Rekayasa Elektronika*, Vol. 6, No. 1, Pp. 73–78, 2023.

[40] A. Widianti And I. Pratama, "Penanganan Missing Values Dan Prediksi Data Timbunan Sampah Berbasis Machine Learning," *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, Vol. 9, No. 2, Pp. 242–251, 2024.

[41] A. G. Lazuardy And H. Setiaji, "Data Cleansing Pada Data Rumah Sakit," *Proceeding Sintak*, Pp. 1–6, 2019.

[42] J. Parhusip, G. Stevans, M. A. Afandy, And M. Rafliansyah, "Analisis Perbandingan Akurasi Algoritma K-Nearest Neighbor (Knn), Decision Tree, Dan Random Forest Dalam Klasifikasi Penyakit Diabetes," Vol. 7, No. 1, Pp. 26–32, 2026.

[43] Agus Ambarwari And Q. J. Adrian, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning Untuk Identifikasi Tanaman," *Resti*, Vol. 1, No. 1, Pp. 19–25, 2020.