INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 11 No. 2 Desember 2025, pp. 433-441

PERANCANGAN SISTEM PREDIKSI RISIKO DIABETES MENGGUNAKAN LOGISTIC REGRESSION DAN RANDOM FOREST

Samuel Paul Jackson S[1], Rafi Abdul Aziz[2], Ichwan Agil Prasetyo[3], M. Raihan Al Ikhsan[4], Abiath Cio[5], Sigit Wibawa[6], Muhammad Muharrom[7]
[1,2,3,4,5,6,7] *Universitas Bina Sarana Informatika, Bekasi, Indonesia*
Responden: saamueeeel11@gmail.com

ABSTRACT
This study aims to design a diabetes risk prediction system using Logistic Regression and Random Forest algorithms, utilizing the Pima Indians Diabetes dataset. The preprocessing stage includes median-based imputation of zero values in medical features and data normalization before splitting into training and testing sets. Both models are trained in parallel, and the best-performing model is selected based on F1-Score to improve the detection accuracy of patients with diabetes (Outcome = 1). Performance evaluation using accuracy, precision, recall, F1-Score, and confusion matrix shows that Random Forest achieves the best results with 74% accuracy, F1-Score of 0.82 for the Non-Diabetes class, and 0.59 for the Diabetes class. The system is integrated into an interactive Gradio interface, allowing users to input medical parameters and obtain real-time risk predictions. The results indicate that the system can support early diabetes detection efficiently, although further improvements are needed to reduce False Negative errors in positive cases.

Keywords: diabetes, risk prediction, Logistic Regression, Random Forest, machine learning

ABSTRAK

Penelitian ini bertujuan merancang sistem prediksi risiko diabetes menggunakan algoritma Logistic Regression dan Random Forest dengan memanfaatkan dataset Pima Indians Diabetes. Tahap pra-pemrosesan meliputi imputasi nilai nol pada fitur medis dan normalisasi data sebelum dibagi menjadi data latih dan uji. Kedua model dilatih secara paralel, dan model terbaik dipilih berdasarkan nilai F1-Score untuk meningkatkan akurasi pendeteksian pasien dengan diabetes (Outcome = 1). Evaluasi performa menggunakan metrik akurasi, presisi, recall, F1-Score, dan confusion matrix menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 74%, F1-Score 0,82 untuk kelas Tidak Diabetes, dan 0,59 untuk kelas Diabetes. Sistem ini diintegrasikan ke antarmuka Gradio interaktif, memungkinkan pengguna memasukkan parameter medis dan memperoleh prediksi risiko secara real-time. Hasil penelitian menunjukkan bahwa sistem mampu mendukung deteksi dini diabetes secara efisien, meskipun perlu peningkatan untuk mengurangi kesalahan False Negative pada pasien positif diabetes.

Kata kunci: diabetes, prediksi risiko, Logistic Regression, Random Forest, machine learning

Riwayat Artikel :

Tanggal diterima : 13-10-2025

Tanggal revisi : 12-11-2025

Tanggal terbit : 01-12-2025

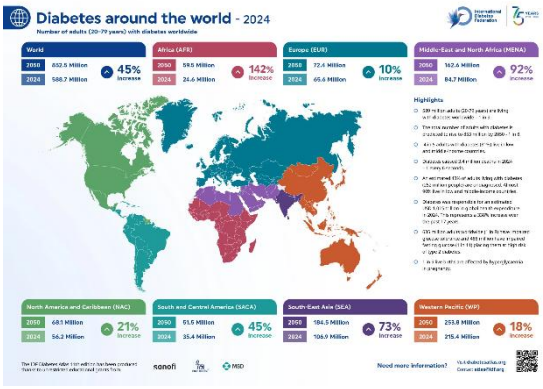
DOI :

https://doi.org/10.31949/infotech.v11i2.16668

1. PENDAHULUAN

Diabetes mellitus merupakan kondisi kronis yang ditandai oleh peningkatan konsentrasi glukosa darah serta munculnya gejala khas berupa urine yang berasa manis dalam jumlah berlebihan. Kelainan utama pada diabetes mellitus terjadi akibat gangguan relatif atau absolut pada hormon insulin, yang mengganggu kemampuan tubuh dalam mengatur kadar glukosa darah secara normal (Decroli, 2022) Patofisiologi utama pada kasus diabetes tipe 2 melibatkan resistensi insulin, yaitu kegagalan jaringan target dalam merespon insulin secara adekuat, sehingga glukosa darah tidak dapat diatur dengan normal (Widiasari et al., 2021). Insulin merupakan satu-satunya hormon yang berfungsi menurunkan kadar glukosa dalam darah, sehingga kekurangan ataupun resistensi terhadap hormon ini menjadi penyebab mendasar terjadinya hiperglikemia (Bilous et al., 2021)

Berdasarkan laporan International Diabetes Federation (IDF) edisi tahun 2025, jumlah penderita diabetes di seluruh dunia mencapai 589 juta orang pada tahun 2024, dan angka tersebut diproyeksikan meningkat hingga 853 juta orang pada tahun 2050 (Ali et al., 2025; Duncan et al., 2025) apabila tren ini terus berlanjut. Peningkatan yang sangat signifikan ini menunjukkan bahwa diabetes merupakan masalah kesehatan global (Guzman-Vilca & Carrillo-Larco, 2025).



Gambar 1. Grafik prevalensi diabetes global berdasarkan IDF Diabetes Atlas 2025.

Deteksi dini terhadap risiko diabetes menjadi semakin penting untuk mencegah komplikasi jangka panjang seperti penyakit jantung, gagal ginjal, hingga kerusakan saraf. Komplikasi yang umum terjadi meliputi gangguan kardiovaskular seperti penyakit jantung koroner, kerusakan ginjal yang bisa berkembang menjadi gagal ginjal kronis, serta gangguan saraf (neuropati) yang dapat menyebabkan luka kaki atau ulkus diabetik(Ardini & Halim, 2023). Selain itu, prevalensi komplikasi seperti nefropati dan penyakit kardiovaskular pada penderita diabetes cukup tinggi, sehingga penting dilakukan deteksi dan pengelolaan sejak dini untuk mencegah penurunan kualitas hidup secara drastis (Paisal et al., 2024). Oleh karena itu, pemanfaatan teknologi komputasi dan machine learning menjadi salah satu pendekatan modern yang mampu meningkatkan efektivitas identifikasi dini terhadap penyakit ini (Ginting et al., 2022).

Machine Learning (ML) merupakan cabang dari kecerdasan buatan (Artificial Intelligence/AI) yang memungkinkan komputer untuk belajar dari data dan meningkatkan kinerjanya seiring waktu, tanpa perlu diprogram secara eksplisit untuk setiap tugas tertentu (Barbierato & Gatti, 2024). ML mengintegrasikan prinsip-prinsip ilmu komputer dan statistik sehingga sistem dapat meningkatkan performanya berdasarkan pengalaman (Trisal & Mandloi, 2021). Dalam praktiknya, ML membutuhkan data berlabel (khususnya pada supervised learning) dan algoritma yang sesuai, agar model dapat mempelajari hubungan antara input dan output, serta melakukan prediksi atau klasifikasi secara (Nurhalizah et al., 2024)

Machine Learning umumnya diklasifikasikan menjadi tiga pendekatan utama: Supervised Learning, Unsupervised Learning, dan Reinforcement Learning. Supervised Learning menggunakan data berlabel, sehingga model dilatih dengan pasangan input-output yang telah diketahui, memungkinkan prediksi atau klasifikasi hasil untuk data baru (misalnya regresi, klasifikasi). Unsupervised Learning bekerja pada data tanpa label, berfokus menemukan pola, struktur, atau kelompok tersembunyi dalam data — cocok untuk eksplorasi data atau clustering. Sementara itu, Reinforcement Learning berbeda karena model belajar melalui interaksi dengan lingkungan: model membuat tindakan, mendapat umpan balik (reward/punishment), dan dari situ belajar strategi terbaik untuk mencapai tujuan.(Ahmad Nur Ihsan Purwanto et al., 2025)

Logistic Regression adalah algoritma machine learning yang digunakan untuk memprediksi probabilitas terjadinya suatu peristiwa, dalam hal ini risiko diabetes (Rassiyanti et al., 2025). Secara dasar, Logistic Regression berasal dari model regresi linier, yang dirumuskan sebagai:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Dimana $y$adalah output prediksi, $X_1, X_2, \ldots, X_n$adalah fitur input, dan $\beta_0, \beta_1, \ldots, \beta_n$adalah koefisien yang dipelajari dari data. Namun, regresi linier tidak cocok untuk memprediksi probabilitas karena hasilnya bisa berada di luar rentang [0,1]. Untuk itu, output linier ini kemudian diubah menjadi probabilitas menggunakan fungsi logit, yaitu transformasi logaritma dari odds (perbandingan probabilitas kejadian terhadap tidak terjadinya kejadian), sehingga terbentuk rumus Logistic Regression:

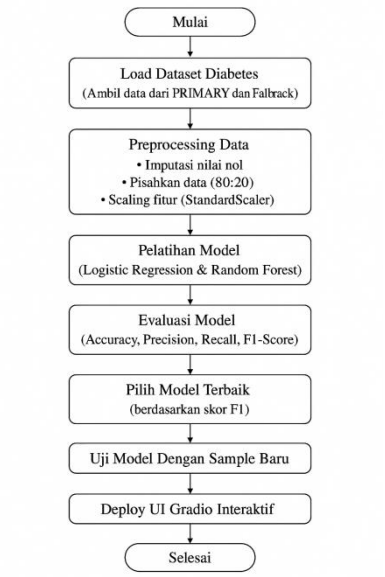
$$logit(p) = in\left(\frac{p}{1-p}\right)\ \beta_0\ +\ \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Dimana $p$adalah probabilitas terdiagnosis diabetes, $X_1, X_2, \ldots, X_n$adalah fitur medis seperti kadar Glukosa, BMI, dan usia, dan $\beta_0, \beta_1, \ldots, \beta_n$adalah koefisien yang dipelajari dari data (Inonu et al., 2025). Transformasi logaritma ini memungkinkan model mengubah output linier menjadi probabilitas antara 0 dan 1, sehingga hasil prediksi dapat diinterpretasikan langsung sebagai risiko pasien. Sementara itu, Random Forest adalah algoritma ensemble yang membangun banyak pohon keputusan secara acak dan menggabungkan hasil prediksi masing-masing pohon melalui voting untuk meningkatkan akurasi dan mengurangi overfitting (Inonu et al., 2025). Meskipun Random Forest tidak menggunakan logaritma dalam perhitungannya, algoritma ini dapat bekerja bersama Logistic Regression dalam sistem prediksi karena kedua metode menangani klasifikasi dari perspektif berbeda: Logistic Regression memberikan prediksi probabilitas berbasis logaritma, sementara Random Forest menangani hubungan non-linear dan interaksi kompleks antar fitur.

Kombinasi kedua algoritma ini memungkinkan sistem secara otomatis memilih model terbaik berdasarkan F1 Score, sehingga prediksi risiko diabetes menjadi lebih akurat. Logistic Regression memberikan probabilitas yang mudah diinterpretasikan, sedangkan Random Forest meningkatkan kemampuan deteksi terhadap pola yang kompleks, sehingga pipeline prediksi menjadi lebih robust dan dapat diandalkan

2. METODE

2.1. Tahapan Penelitian



Gambar 2. Tahapan Penelitian

Tahapan penelitian ini dilakukan dengan beberapa langkah utama yang ditunjukkan di gambar 2.

a. **Load Dataset Diabetes**
Tahap ini dilakukan dengan memuat dataset diabetes dari dua sumber berbeda. Sistem pertama-tama mencoba mengunduh data dari sumber utama (PRIMARY), dan apabila terjadi kegagalan, proses secara otomatis beralih memuat data dari sumber cadangan (Fallback). Langkah ini bertujuan memastikan bahwa dataset tetap tersedia dan penelitian dapat berjalan tanpa hambatan.*Preprocessing Data*
Setelah dataset dimuat oleh *Python*, dataset akan dirubah kedalam bentuk matriks agar dapat dipahami oleh model *Machine Learning*. Setelah dirubah, dataset akan di bagi menjadi dua bagian yaitu 80% untuk *training* dan 20% untuk *testing* dengan menggunakan *train_test_split*

b. **Preprocessing Data**
Pada tahap preprocessing, data dibersihkan dan dipersiapkan agar layak digunakan oleh model. Imputasi nilai nol dilakukan untuk mengganti nilai 0 pada kolom medis yang tidak logis, seperti tekanan darah atau kadar glukosa. Setelah itu, data dipisahkan menjadi data latih dan data uji dengan komposisi 80% untuk pelatihan dan 20% untuk pengujian. Selanjutnya, seluruh fitur dinormalisasi menggunakan StandardScaler agar model dapat bekerja secara optimal dengan skala data yang seragam.*split* tadi.

c. **Pelatihan Model (Logistic Regression & Random Forest)**
Dua algoritma machine learning kemudian dilatih menggunakan data hasil preprocessing. Logistic Regression dipakai sebagai model dasar yang bersifat linear, sedangkan Random Forest digunakan sebagai model yang lebih kompleks berbasis ensemble. Kedua model ini dilatih menggunakan data latih yang telah dinormalisasi.

d. **Evaluasi Model (Accuracy, Precision, Recall, F1-Score)**
Setelah pelatihan selesai, kedua model dievaluasi menggunakan data uji. Evaluasi dilakukan dengan beberapa metrik, yaitu accuracy, precision, recall, dan F1-score. Metrik-metrik ini digunakan untuk menilai sejauh mana model mampu melakukan prediksi dengan benar, terutama pada kasus klasifikasi yang sensitif seperti deteksi diabete.

e. **Pilih Model Terbaik (berdasarkan skor F1)**
Dari hasil evaluasi, model terbaik dipilih berdasarkan nilai F1-score tertinggi. F1-score dijadikan prioritas karena lebih akurat dalam menangani ketidakseimbangan data dan menilai kemampuan model dalam mendeteksi pasien yang terindikasi diabetes.

f. **Uji Model Dengan Sample Baru**
Model yang terpilih kemudian diuji menggunakan sampel baru untuk memastikan bahwa model mampu memberikan prediksi

yang stabil dan tetap konsisten saat menerima data yang belum pernah dilihat sebelumnya

g. Deploy UI Gradio Interaktif

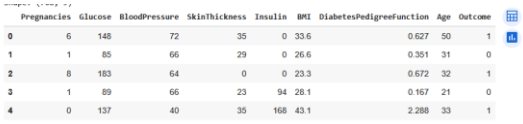
Tahap terakhir adalah melakukan deployment model ke dalam sebuah antarmuka interaktif menggunakan Gradio. Antarmuka ini memungkinkan pengguna memasukkan nilai-nilai input seperti kadar glukosa, tekanan darah, dan usia untuk mendapatkan hasil prediksi secara langsung dan mudah dipahami

2.2. Data Penelitian

a. **Sumber Data**

Data dari penelitian ini bersumber dari platform *Kaggle* dengan judul "Pima Indians DiabetesDatabase" https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Dataset Pima Indians Diabetes berisi 768 pasien dengan 9 kolom, termasuk fitur medis seperti Glucose, BloodPressure, BMI, Age, dan kolom target Outcome (0 = tidak diabetes, 1 = diabetes). Beberapa nilai nol diimputasi dengan median agar model machine learning seperti Logistic Regression dan Random Forest bisa memprediksi risiko diabetes dengan akura



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Gambar 3. Distribusi Kelas Dalam File "Dataset Pima Indians Diabetes"

Dari gambar tersebut dapat disimpulkan bahwa dataset berisi sejumlah fitur medis seperti jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, BMI, fungsi riwayat keluarga diabetes, dan usia. Setiap baris mewakili satu data pasien, sedangkan kolom *Outcome* menunjukkan label kondisi pasien, yaitu 1 untuk pasien yang terindikasi diabetes dan 0 untuk pasien yang tidak terindikasi diabetes. Berdasarkan contoh yang terlihat, terdapat kombinasi berbagai nilai medis yang akan digunakan untuk melatih model dalam mendeteksi risiko diabetes.

b. **Bentuk Data**

Dataset berbentuk *CSV (Comma Seperated Value)* yang dapat dibuka dengan *Microsoft Excel* .

2.3. Preprocessing Data

Sebelum digunakan oleh model machine learning, data diabetes perlu diproses agar dapat dibaca sebagai matriks numerik. Dataset Pima Indians Diabetes di-load menggunakan pandas dan memuat fitur medis seperti *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age*, serta kolom target *Outcome*. Nilai nol pada beberapa kolom diimputasi dengan median untuk menjaga konsistensi dan mengurangi bias.

Data kemudian dipisahkan menjadi fitur (X) dan label (y), dan dibagi menjadi data latih (80%) dan data uji (20%) menggunakan train_test_split. Semua fitur diskalakan dengan StandardScaler agar memiliki skala yang sama, sehingga model Logistic Regression dan Random Forest dapat bekerja optimal.

Tahap preprocessing ini memastikan data siap dilatih. Logistic Regression memprediksi probabilitas diabetes melalui transformasi logit (logaritma dari odds), sedangkan Random Forest membangun banyak pohon keputusan dan menggabungkan prediksi untuk meningkatkan akurasi. Kombinasi kedua algoritma ini memungkinkan sistem memilih model terbaik secara otomatis berdasarkan F1 Score, sehingga prediksi risiko diabetes lebih akurat

Contoh kode *Python* dapat dilihat di Tabel 1

Tabel 1. Contoh Kode *Python* Untuk *Preprocessing* Data.

Kode	Penjelasan
df = pd.read_csv(PRIMARY) assert set(COLS) <= set(df.columns) print("Load dari PRIMARY mirror")	Kode ini membaca file CSV dari URL dan menyimpannya dalam DataFrame., memastikan semua kolom yang dibutuhkan (COLS) ada di dataset
X = df['Teks'] y = df['label']	Kode ini digunakan untuk memisah tabel berdasarkan teks (X) dan label (y).
X = df.drop ("Outcome", axis=1) y = df ["Outcome"].astype(int)	Memisahkan data menjadi fitur (X) dan target (y). X berisi semua kolom kecuali Outcome, sedangkan y adalah kolom Outcome yang ingin diprediksi..

Algoritma yang digunakan untuk melatih model adalah Logistic Regression dan Random Forest. Logistic Regression memprediksi probabilitas risiko diabetes berdasarkan hubungan linier antar fitur yang ditransformasikan dengan fungsi logit, sedangkan Random Forest membangun banyak pohon keputusan secara acak dan menggabungkan hasil prediksi untuk meningkatkan akurasi dan

mengurangi overfitting. Contoh kode Python dapat dilihat di Tabel 2.

Tabel 2. Contoh Kode *Python* Untuk Preprocessing Data.

Kode	Penjelasan
models = { "LogisticRegression": LogisticRegression(max_iter=200, random_state=42), "RandomForest": RandomForestClassifier(n_estimators=200, random_state=42) }	Mendefinisikan dua algoritma Machine Learning yang akan digunakan: Logistic Regression untuk prediksi probabilitas dan Random Forest sebagai ensemble decision tree
mdl.fit(X_train_s, y_train)	Melatih model (fit) menggunakan data training yang telah diskalakan (X_train_s) dan label (y_train)..

2.4. Evaluasi Model (*Accuracy, Precision, Recall, F1-Score)*

Evaluasi model dilakukan untuk menilai kemampuan algoritma dalam memprediksi risiko diabetes pada pasien berdasarkan data medis. Dua algoritma yang diuji adalah **Logistic Regression** dan **Random Forest**, kemudian model dengan **F1-Score tertinggi** dipilih sebagai model utama.:

a. Akurasi (*Accuracy*): Akurasi mengukur proporsi prediksi yang benar dari seluruh prediksi (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

b. *Precision*: Precision mengukur proporsi prediksi positif yang benar (3).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

c. *Recall*: Recall mengukur kemampuan model mendeteksi semua kasus positif: (4).

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

d. F1-Score: Menunjukkan rata-rata harmonik dari *precision* dan *recall* yang dapat memberikan gambaran seimbang antara kedua metrik tersebut. Rumus *F1-Score* dapat dilihat di rumus (5).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Keterangan:

TP (True Positive): pasien benar-benar diabetes dan diprediksi positif
TN (True Negative): pasien tidak diabetes dan diprediksi negatif
FP (False Positive): pasien tidak diabetes tapi diprediksi positif
FN (False Negative): pasien diabetes tapi diprediksi negatif
.

2.5. Pengujian Model

Setelah model dilatih, akan dilakukan pengujian terhadap beberapa baru diluar dataset untuk mengetahui apakah model dapat memprediksi diabetes dengan benar atau tidak. Contoh kode *Python* dapat dilihat di Tabel 3.

Tabel 3. Contoh Kode *Python* Untuk Pengujian Model.

Kode	Penjelasan
python patient_samples = [[2, 120, 70, 20, 80, 25.0, 0.5, 30], [4, 150, 85, 30, 100, 32.0, 0.7, 45], [1, 90, 60, 15, 50, 20.0, 0.3, 25]] 	Membuat array sampel pasien baru yang akan diuji model. Setiap elemen array mewakili satu pasien dengan fitur: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
python patient_samples_s = scaler.transform(patient_samples) predictions = best_model.predict(patient_samples_s) proba = best_model.predict_proba(patient_samples_s)[:,1] 	Data pasien distandarisasi sesuai training, kemudian diprediksi kelas (diabetes/tidak) dan probabilitas kelas positif.model.

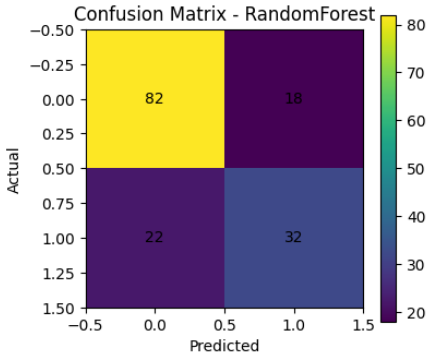
3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset Pima Indians Diabetes untuk memprediksi risiko diabetes berdasarkan parameter medis. Data diproses dengan imputasi nilai nol dan dibagi menjadi 80% untuk training dan 20% untuk testing. Dua algoritma, yaitu Logistic Regression dan Random Forest, diuji, dan model dengan F1-Score tertinggi dipilih sebagai model terbaik. Evaluasi performa dilakukan menggunakan metrik accuracy, precision, recall, dan F1-Score, serta divisualisasikan melalui Confusion Matrix. Model terpilih dapat digunakan untuk memprediksi risiko pasien baru melalui antarmuka Gradio, sementara Random Forest memungkinkan analisis feature importance untuk mengidentifikasi

variabel medis yang paling berpengaruh terhadap prediksi..

3.1. Hasil *Confusion Matrix Naïve Bayes*

Confusion Matrix merupakan metode evaluasi yang umum digunakan untuk menilai performa model klasifikasi. Matriks ini menyajikan perbandingan antara label sebenarnya dengan hasil prediksi model, sehingga memberikan gambaran yang lebih mendalam mengenai kemampuan model dalam mengenali setiap kelas. Menurut Kelleher dan Tierney (2018), *confusion matrix* sangat penting karena membantu mengidentifikasi pola kesalahan yang tidak terlihat hanya dari nilai akurasi saja..



Gambar 4. Hasil Confusion Matrix Algoritma RandomForest

Dari Gambar 4 dapat disimpulkan bahwa:

a. Model berhasil menklasifikasikan dengan benar 82 data yang tidak diabetes (kelas 0) dari total data kelas 0, sehingga mayoritas pasien yang sehat dapat dikenali dengan baik oleh model.
b. Model juga berhasil menklasifikasikan dengan benar 32 data yang terindikasi diabetes (kelas 1) dari total data kelas 1. Ini menunjukkan bahwa model cukup mampu mendeteksi pasien yang memiliki diabetes.
c. Kesalahan terbesar terjadi ketika model salah mengklasifikasikan 22 pasien yang sebenarnya diabetes namun diprediksi sebagai tidak diabetes (False Negative). Hal ini cukup penting karena kasus FN bisa berbahaya: pasien sebenarnya positif diabetes tetapi tidak terdeteksi.

Selain itu, terdapat 18 kasus False Positive, yaitu pasien yang sebenarnya tidak diabetes tetapi diprediksi sebagai diabetes. Meskipun tidak berbahaya seperti FN, hal ini tetap menurunkan akurasi model.

Secara keseluruhan, model RandomForest yang digunakan mencapai performa yang baik, mampu mengenali pola dataset dengan cukup tepat. Namun, nilai FN yang cukup tinggi menunjukkan bahwa model masih perlu ditingkatkan pada kemampuan mendeteksi pasien yang benar-benar terindikasi diabetes.

3.2. *Accuracy, Precision, Recall, dan F1-Score model*

Evaluasi model dilakukan di environment Python untuk memperoleh metrik kinerja seperti *accuracy*, *precision, recall, dan F1-score*. Hasil evaluasi tersebut ditampilkan pada Classification Report pada (Gambar 5.)

```
--- Classification Report (model terpilih) ---
              precision    recall  f1-score   support

           0       0.79      0.82      0.80       100
           1       0.64      0.59      0.62        54

    accuracy                           0.74       154
   macro avg       0.71      0.71      0.71       154
weighted avg       0.74      0.74      0.74       154
```

Gambar 5. Laporan Hasil Klasifikasi Model di Python

Perhitungan akurasi model dapat dijelaskan melalui hasil Confusion Matrix yang diperoleh dari model RandomForest. Berdasarkan matriks tersebut, terdapat 82 prediksi benar pada kelas 0 (True Negative) dan 32 prediksi benar pada kelas 1 (True Positive), sehingga total prediksi yang tepat adalah 114 data. Sementara itu, jumlah keseluruhan data uji adalah 154 yang terdiri dari 100 data kelas 0 dan 54 data kelas 1. Dengan menggunakan rumus akurasi, yaitu total prediksi benar dibagi dengan total data uji, diperoleh nilai 114/154 = 0.74026. Dengan demikian, model memiliki akurasi sebesar 0.74 atau 74%, yang sesuai dengan nilai akurasi yang ditampilkan pada laporan klasifikasi Python.

Namun demikian, model menunjukkan nilai recall yang lebih rendah pada kelas Diabetes (kelas 1). Hal ini dapat disebabkan oleh ketidakseimbangan kelas, di mana jumlah sampel Non-Diabetes (100 data) lebih besar dibandingkan sampel Diabetes (54 data). Kondisi ini membuat model cenderung lebih "bias" terhadap kelas mayoritas sehingga kesalahan dalam mendeteksi pasien yang benar-benar Diabetes (False Negative) menjadi lebih tinggi.

Untuk mengatasi ketidakseimbangan tersebut, sebenarnya terdapat beberapa teknik yang umum digunakan, seperti SMOTE (*Synthetic Minority Over-sampling Technique)* dan *class weighting*. Pada penelitian ini, pendekatan yang digunakan adalah class weighting, yaitu memberikan bobot lebih besar pada kelas minoritas saat proses pelatihan model. Metode ini tidak mengubah jumlah data tetapi membantu model untuk lebih memperhatikan kesalahan pada kelas Diabetes. Meskipun demikian, hasil evaluasi tetap menunjukkan bahwa performa pada kelas Diabetes masih dapat ditingkatkan, sehingga teknik resampling seperti SMOTE dapat dipertimbangkan pada penelitian selanjutnya.

a. **Akurasi *(Accuracy)***

   Perhitungan akurasi juga dapat dijelaskan menggunakan Confusion Matrix hasil model.

Dari confusion matrix model RandomForest (82, 18, 22, 32), jumlah prediksi yang benar adalah:

82 (True Negatif) +32 (True Positive) =114

Total data uji:

100 (kelas 0) + 54 (kelas 1)=154

Maka akurasi model dihitung sebagai:

$$\text{Accuracy} = \frac{114}{154} = 0.74026$$

Dengan demikian, model memiliki akurasi sekitar 0.74 (74%), sesuai dengan nilai yang ditampilkan pada laporan klasifikasi Python.

b. ***Precision***

Perhitungan *precision* dapat dilakukan dengan menggunakan rumus. Berdasarkan hasil *Confusion Matrix* yang *True Positive* dan *False Positive* masing-masing kelas sebagai berikut (Table 4).

Tabel 4. Nilai *True Positive* dan *False Positive* Setiap Kelas

Kelas	TP	FP
(0) (Tidak Diabetes)	82	18
(1) (Diabetes)	32	22

Dari Tabel 4 dapat dihitung nilai *precision* setiap kelas:
0 (Tidak Diabetes) :

$$Precision = \frac{82}{82 + 18}$$

$$Precision = \frac{82}{100}$$

$$Precision = 0.82$$

1 (Diabetes) :

$$Precision = \frac{32}{32 + 22}$$

$$Precision = \frac{32}{54}$$

$$Precision = 0.5926 = 0.59$$

Tabel 5. Hasil Perhitungan *Precision* Setiap Kelas

Kelas	Nilai Precision
(0) (Tidak Diabetes)	0.82 (82%)
(1) (Diabetes)	0.59 (59%)

Berdasarkan Tabel 5, kelas 0 memiliki precision lebih tinggi dibandingkan kelas 1. Hal ini menunjukkan bahwa model lebih tepat dalam mengidentifikasi pasien yang tidak memiliki diabetes, namun masih kurang presisi dalam memprediksi pasien yang benar-benar memiliki diabetes.

c. ***Recall***

Perhitungan *recall* untuk masing-masing kelas menggunakan rumus Dengan menggunakan hasil *Confusion Matrix* di Gambar 4, dapat diketahui bahwa nilai *True Positive* dan *False Negative* masing-masing kelas sesuai dengan (Tabel 6).

Tabel 6. Nilai *True Positive* dan *False Negative* Setiap Kelas

Kelas	TP	FP
(0) (Tidak Diabetes)	82	18
(1) (Diabetes)	32	22

Dari Tabel 6, dapat dihitung nilai *recall* setiap kelas sebagai berikut:

0 (Tidak Diabetes) :

$$Recall = \frac{82}{82 + 18}$$

$$Recall = \frac{82}{100}$$

$$Recall = 0.82$$

1 (Diabetes):

$$Recall = \frac{32}{32 + 22}$$

$$Recall = \frac{32}{54}$$

$$Recall = 0.59$$

Tabel 7. Hasil Perhitungan *Recall* Setiap Kelas

Kode	Nilai Recall
(0) (Tidak Diabetes)	0.82 (82%)
(1) (Diabetes)	0.59 (59%)

Dari Tabel 7, Nilai recall yang rendah pada kelas 1 (Diabetes) menunjukkan bahwa model masih sering gagal mendeteksi pasien yang sebenarnya positif diabetes. Hal ini ditandai dengan jumlah False Negative yang cukup tinggi, terutama karena adanya

ketidakseimbangan jumlah data antara kelas Non-Diabetes dan Diabetes.

d. ***F1-Score***

Nilai *F1-Score* masing-masing kelas dapat dihitung menggunakan rumus. Dengan nilai *precision* dan *recall* yang telah dihitung sebelumnya, kita dapat menghitung hasil *F1-Score* masing-masing kelas sebagai berikut.

F1-score Kelas 0

$$F1 = 2 \times \frac{0.82 \times 0.82}{0.82 + 0.82}$$

$$F1 = 0.82$$

F1-score Kelas 1

$$F1 = 2 \times \frac{0.59 \times 0.59}{0.59 + 0.59}$$

$$F1 = 0.59$$

Tabel 8. Hasil Perhitungan *F1-Score* Setiap Kelas

Kode	Nilai F1-Score
(0) Tidak Diabetes	0.82(82%)
(1) Diabetes	0.58(59%)

Dari Tabel 8 F1-score kelas 1 lebih rendah, yang mengindikasikan bahwa keseimbangan antara precision dan recall masih belum optimal untuk mendeteksi diabetes.

Berdasarkan hasil evaluasi menggunakan Confusion Matrix dan metrik kinerja model RandomForest, diperoleh akurasi sebesar 74%, yang menunjukkan bahwa model memiliki kemampuan prediksi yang cukup baik. Namun, performa model pada masing-masing kelas menunjukkan ketidakseimbangan.

Model bekerja sangat baik pada kelas Tidak Diabetes, dengan precision, recall, dan F1-score masing-masing sebesar 0.82, yang menandakan kemampuan tinggi dalam mengidentifikasi pasien yang tidak memiliki diabetes. Sebaliknya, performa pada kelas Diabetes masih rendah dengan precision, recall, dan F1-score masing-masing 0.59, sehingga model masih sering gagal mendeteksi pasien yang sebenarnya positif diabetes.

Secara keseluruhan, model lebih akurat untuk mengenali kondisi Tidak Diabetes, tetapi masih kurang sensitif dalam mendeteksi Diabetes. Oleh karena itu, peningkatan model diperlukan, terutama untuk mengurangi kasus False Negative agar deteksi diabetes menjadi lebih optimal.

4. KESIMPULAN

Penelitian ini berhasil merancang sistem prediksi risiko diabetes dengan memanfaatkan algoritma Logistic Regression dan Random Forest menggunakan dataset Pima Indians Diabetes. Tahapan pra-pemrosesan yang meliputi imputasi nilai nol dan normalisasi data terbukti penting dalam meningkatkan kualitas data sebelum pelatihan model. Dari dua algoritma yang diuji, Random Forest menunjukkan performa terbaik berdasarkan nilai F1-Score dan dipilih sebagai model utama.

Hasil evaluasi menunjukkan bahwa model memiliki akurasi sebesar 74%, dengan kinerja yang sangat baik pada kelas Tidak Diabetes (precision, recall, dan F1-score = 0.82). Namun, performa model menurun pada kelas Diabetes, yang ditunjukkan oleh nilai precision, recall, dan F1-score sebesar 0.59. Hal ini mengindikasikan bahwa model masih sering mengalami kesalahan False Negative, yaitu gagal mendeteksi pasien yang sebenarnya positif diabetes. Kondisi ini menjadi perhatian penting karena dapat berdampak pada keterlambatan penanganan medis.

Secara keseluruhan, sistem prediksi yang dikembangkan mampu memberikan estimasi risiko diabetes secara cukup akurat dan dapat digunakan untuk mendukung deteksi dini melalui antarmuka Gradio yang interaktif dan mudah digunakan. Meski demikian, peningkatan model tetap diperlukan, terutama dalam memperkuat kemampuan mendeteksi kelas positif dengan mengurangi False Negative melalui teknik seperti penyeimbangan data, tuning hyperparameter, atau penggunaan algoritma alternatif

PUSTAK

Ahmad Nur Ihsan Purwanto, Muhammad Naufal Ammr Dzakwan, & Fadillah Dani Prawoto. (2025). Tren dan Perkembagan Supervised versus Unsupervised Learning. *Jurnal Teknik Informatika Dan Teknologi Informasi*, *5*(2), 619–625. https://doi.org/10.55606/jutiti.v5i2.5742

Ali, A. A., Galal, G. R., & Hassan, H. S. (2025). Diabetes Prediction on Pima Indian Dataset Using Machine Learning Techniques. *International Journal of Environmental Sciences*, 529–550. https://doi.org/10.64252/3a8wqx36

Ardini, F., & Halim, S. (2023). HUBUNGAN HBA1C DENGAN KOMPLIKASI MAKROVASKULAR PADA DMT2 DI RS HERMINA KEMAYORAN 2022. *Jurnal Kesehatan Tambusai*, *4*(4), 6772–6778. https://doi.org/10.31004/jkt.v4i4.22366

Barbierato, E., & Gatti, A. (2024). The Challenges of Machine Learning: A Critical Review. *Electronics*, *13*(2), 416. https://doi.org/10.3390/electronics13020416

Bilous, R., Donnelly, R., & Idris, I. (2021). *Handbook of Diabetes*. John Wiley & Sons.

Decroli, E. (2022). Mekanisme Molekuler Dari Resistensi Insulin Pada Diabetes Melitus Tipe Dua. *Majalah Kedokteran Andalas*, *45*(4), 610–618. https://doi.org/10.25077/mka.v45.i4.p610-618.2022

Duncan, B. B., Magliano, D. J., & Boyko, E. J. (2025). IDF Diabetes Atlas 11th edition 2025: global prevalence and projections for 2050. *Nephrology Dialysis Transplantation*. https://doi.org/10.1093/ndt/gfaf177

Ginting, J., Ginting, R., & Hartono, H. (2022). DETEKSI DAN PREDIKSI PENYAKIT DIABETES MELITUS TIPE 2 MENGGUNAKAN MACHINE LEARNING (SCOOPING REVIEW). *Jurnal Keperawatan Priority*, *5*(2), 93–105. https://doi.org/10.34012/jukep.v5i2.2671

Guzman-Vilca, W. C., & Carrillo-Larco, R. M. (2025). Number of People with Type 2 Diabetes Mellitus in 2035 and 2050: A Modelling Study in 188 Countries. *Current Diabetes Reviews*, *21*(1). https://doi.org/10.2174/0115733998274323231230131843

Inonu, O. Y., Magda, K., & Amarudin, A. (2025). Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *15*(1), 1. https://doi.org/10.36448/expert.v15i1.4312

Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 61–72. https://doi.org/10.54082/jiki.168

Paisal, P., Arifin, A. Y., & Primasari, P. (2024). Komplikasi Kardiovaskuler dan Ginjal Pasien Diabetes Melitus di Rumah Sakit Rujukan. *PROSIDING KONFERENSI NASIONAL ILMU KESEHATAN STIKES ADI HUSADA 2023*, *2*(1), 33. https://doi.org/10.37036/prosiding.v2i1.601

Rassiyanti, L., Farid, F., & Pitri, R. (2025). Diabetes risk prediction using logistic regression model. *Desimal: Jurnal Matematika*, *8*(1), 41–50. https://doi.org/10.24042/djm.v8i1.26493

Trisal, A., & Mandloi, D. (2021). MACHINE LEARNING: AN OVERVIEW. *International Journal of Research -GRANTHAALAYAH*, *9*(7), 343–348. https://doi.org/10.29121/granthaalayah.v9.i7.2021.4120

Widiasari, K. R., Wijaya, I. M. K., & Suputra, P. A. (2021). DIABETES MELITUS TIPE 2: FAKTOR RISIKO, DIAGNOSIS, DAN TATALAKSANA. *Ganesha Medicine*, *1*(2), 114. https://doi.org/10.23887/gm.v1i2.40006