Laporan Proyek Machine Learning

Ai Nur Azizah – 2306077
Muhammad Agus solehhudin – 2306033

IMPLEMENTASI PREDIKSI DIABETES MELITUS MENGGUNAKAN RANDOM FOREST PADA DATASET PIMA INDIANS

Diabetes mellitus merupakan salah satu penyakit kronis paling mematikan yang saat ini menjadi masalah kesehatan global dengan prevalensi yang terus meningkat dari tahun ke tahun. Data dari World Health Organization (WHO) memperkirakan bahwa pada tahun 2021 terdapat sekitar 422 juta orang dewasa hidup dengan diabetes di seluruh dunia, dan angka ini diprediksi terus bertambah seiring dengan penuaan populasi, perubahan gaya hidup, dan peningkatan tingkat obesitas (Noviyanti & Alamsyah, 2024). Jika tidak ditangani dengan tepat, diabetes dapat menyebabkan berbagai komplikasi serius, mulai dari penyakit jantung, gagal ginjal, neuropati, hingga kematian (Zhang, 2025). Oleh karena itu, upaya deteksi dini dan pencegahan diabetes menjadi sangat vital untuk mengurangi beban kesehatan masyarakat secara global.

Metode konvensional dalam memprediksi risiko diabetes, seperti regresi logistik, sering kali gagal menangkap kompleksitas hubungan nonlinier antar variabel klinis dan gaya hidup yang mempengaruhi terjadinya diabetes tipe 2. Di era modern, perkembangan machine learning (ML) telah menghasilkan algoritma prediksi yang mampu mengidentifikasi pola tersembunyi dalam data dan memberikan akurasi lebih tinggi dibandingkan pendekatan tradisional (Massari et al., 2024). Berbagai penelitian telah membuktikan bahwa algoritma seperti Random Forest dan XGBoost mampu mengungguli metode konvensional dalam memodelkan risiko diabetes tipe 2, khususnya dalam menangani data dengan interaksi fitur yang kompleks dan isu imbalance kelas (Massari et al., 2024). Misalnya, studi terbaru menunjukkan XGBoost memiliki akurasi, sensitivitas, dan AUC-ROC tertinggi untuk prediksi diabetes, sementara Random Forest juga menunjukkan kinerja yang konsisten tinggi dengan tingkat akurasi hingga 87% (Zhang, 2025).

Implementasi machine learning dalam deteksi dini diabetes juga berpotensi besar mendukung pengambilan keputusan klinis berbasis data. Melalui analisis feature importance, fitur-fitur seperti kadar glukosa, indeks massa tubuh (BMI), dan usia teridentifikasi sebagai prediktor paling signifikan, yang konsisten dengan literatur klinis internasional. Keunggulan ini membuka peluang untuk integrasi sistem prediksi otomatis di layanan kesehatan primer, memudahkan penyaringan populasi berisiko tinggi dengan cara yang efisien, akurat, dan hemat biaya. Dengan demikian, optimalisasi penggunaan algoritma machine learning sangat penting dalam mendukung upaya pencegahan dan penanganan diabetes di masa depan, sekaligus menekan angka komplikasi serta beban ekonomi akibat penyakit ini.

.

Referensi

- Massari, H. El, Gherabi, N., Qanouni, F., & Mhammedi, S. (2024). Diabetes Prediction Using Machine Learning with Feature Engineering and Hyperparameter Tuning. *International Journal of Advanced Computer Science and Applications*, *15*(8), 171–179. https://doi.org/10.14569/IJACSA.2024.0150818
- Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. *Journal of Information System Exploration and Research*, *2*(1), 41–48. https://doi.org/10.52465/joiser.v2i1.245
- Zhang, Z. (2025). *Comparison of Machine Learning Models for Predicting Type 2 Diabetes Risk Using the Pima Indians Diabetes Dataset*. *4*(1), 65–71. https://doi.org/10.56397/JIMR/2025.02.07

Bussiness Understanding

Problem Statements

1. Mengapa deteksi dini risiko diabetes masih menjadi tantangan di layanan kesehatan saat ini?
   Banyak kasus diabetes yang baru terdiagnosis setelah muncul komplikasi serius, karena keterbatasan metode skrining konvensional yang lambat dan kurang sensitif dalam menangkap risiko secara dini.
2. Mengapa metode prediksi tradisional kurang efektif untuk memetakan risiko diabetes?
   Pendekatan konvensional seperti regresi logistik sering gagal menangkap hubungan nonlinier dan interaksi kompleks antar variabel data kesehatan, sehingga hasil prediksi menjadi kurang akurat dan tidak mampu menyesuaikan dengan karakteristik populasi nyata.
3. Mengapa akurasi model prediksi diabetes berbasis machine learning masih belum optimal dalam praktik nyata?
   Walaupun algoritma machine learning terbukti lebih unggul, banyak studi belum mengoptimalkan pemilihan fitur, penanganan data imbalance, serta hyperparameter tuning secara komprehensif—sehingga model sering kali belum siap diimplementasikan di layanan klinis atau aplikasi screening berbasis masyarakat.
4. Apa yang menyebabkan solusi prediksi risiko diabetes sulit diintegrasikan ke praktik klinis atau sistem kesehatan masyarakat?
   Kebanyakan model yang dikembangkan masih terbatas pada uji coba dataset standar, kurang memperhatikan interpretabilitas, kemudahan penggunaan, dan efisiensi komputasi—padahal aspek-aspek ini krusial agar model benar-benar diterima dan diadopsi oleh tenaga kesehatan.

Goals

1. Mengembangkan model prediktif untuk risiko diabetes menggunakan data kesehatan pasien dengan algoritma Random Forest.
2. Mengevaluasi performa model Random Forest dalam memprediksi risiko diabetes pada data kesehatan.

3. Meningkatkan akurasi dan performa model melalui teknik optimasi seperti hyperparameter tuning.

Solusi Statements

Untuk mencapai tujuan proyek, langkah-langkah berikut akan dilakukan :

1. Eksplorasi dan Pemahaman Data (EDA)
   Menganalisis pola distribusi, hubungan antar variabel, dan potensi outlier dengan dukungan visualisasi data.
2. Pra-pemrosesan Data
   Melakukan pembersihan data, imputasi nilai kosong, serta normalisasi fitur untuk memastikan kualitas data yang optimal.
3. Pembangunan Model Machine Learning
   Mengembangkan model prediksi risiko diabetes menggunakan algoritma Random Forest berbasis data kesehatan pasien.
4. Hyperparameter Tuning
   Melakukan tuning parameter pada model untuk meningkatkan akurasi dan performa prediksi.
5. Evaluasi Performa Model
   Mengevaluasi model dengan metrik Akurasi, Precision, Recall, dan F1-Score guna memastikan keandalan prediksi.

Solusi ini diharapkan dapat menghasilkan model prediksi yang akurat untuk mendukung proses skrining risiko diabetes secara luas.

Data Understanding

Deskripsi Dataset

Dataset yang dipakai adalah Pima Indians Diabetes Dataset yang mencakup 768 data pasien wanita berusia di atas 21 tahun, dengan delapan fitur kesehatan utama termasuk jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh (BMI), riwayat diabetes genetik, dan usia. Kolom target (Outcome) mencerminkan kondisi diabetes pasien (0 = tanpa diabetes, 1 = diabetes), dengan sebaran 500 pasien tanpa diabetes dan 268 pasien dengan diabetes. Dataset ini biasa digunakan sebagai tolok ukur dalam penelitian prediksi risiko diabetes dengan machine learning karena komprehensif dan mencerminkan studi klinis.

Informasi Dataset

Informasi dtaset diberikan menggunakan data.info(). Berikut adalah hasilnya :

Kolom	Tipe Data	Jumlah Data	Deskripsi
Pregnancies	int64	768	Jumlah kehamilan yang pernah dialami pasien

Glucose	float64	768	Konsentrasi glukosa plasma (tes toleransi glukosa)
BloodPressure	float64	768	Tekanan darah diastolik (mm Hg)
SkinThickness	float64	768	Ketebalan lipatan kulit trisep (mm)
Insulin	float64	768	Kadar insulin serum dua jam setelah tes (μU/ml)
BMI	float64	768	Indeks Massa Tubuh (kg/m²)
DiabetesPedigreeFunction	float64	768	Indeks riwayat genetik diabetes dalam keluarga
Age	int64	768	Usia pasien (tahun)
Outcome	int64	768	Target prediksi (0 = Tidak diabetes, 1 = Diabetes)

Dataset ini tidak memiliki missing value pada seluruh kolom, sehingga dapat langsung digunakan untuk proses analisis dan pemodelan machine learning.

\Statistik Deskriptif

Berikut adalah statistik deskriptif untuk fitur numerik dalam dataset:

Fitur	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Pregnancies	768	3.85	3.37	0	1	3	6	17
Glucose	768	120.89	31.97	0	99	117	140.25	199
BloodPressure	768	69.11	19.36	0	62	72	80	122
SkinThickness	768	20.54	15.95	0	0	23	32	99

Fitur	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Insulin	768	79.80	115.24	0	0	30.5	127.25	846
BMI	768	31.99	7.88	0	27.3	32	36.6	67.1
DiabetesPedigreeFunction	768	0.472	0.331	0.078	0.244	0.3725	0.626	2.42
Age	768	33.24	11.76	21	24	29	41	81
Outcome	768	0.349	0.477	0	0	0	1	1

Exploratory Data Analysis (EDA)

1. Distribusi Usia

Rentang usia pasien dalam dataset ini adalah 21 hingga 81 tahun, dengan mayoritas pasien berusia antara 25 hingga 45 tahun. Distribusi usia cenderung normal dengan sedikit skew ke kanan (lebih banyak pasien usia dewasa muda hingga paruh baya).

2. Distribusi BMI

Sebagian besar pasien memiliki BMI pada kisaran 27–37 kg/m², menunjukkan kecenderungan overweight dan obesitas pada populasi ini. Terdapat beberapa nilai BMI yang cukup tinggi sebagai outlier, namun distribusi BMI relatif normal.

3. Hubungan Fitur dengan Diabetes

- **Glukosa & Outcome:** Penderita diabetes cenderung memiliki kadar glukosa lebih tinggi daripada yang tidak diabetes, terlihat jelas pada visualisasi distribusi dan heatmap korelasi.
- **BMI & Outcome:** Pasien dengan diabetes cenderung memiliki BMI lebih tinggi.
- **Umur & Outcome:** Risiko diabetes meningkat pada usia yang lebih tua.
- **Fitur lain:** Kolom seperti Pregnancies dan DiabetesPedigreeFunction juga menunjukkan korelasi positif dengan outcome, sedangkan tekanan darah dan insulin tidak menunjukkan pola sekuat glukosa/BMI.

Kesimpulan Data Understanding

- Dataset sudah bersih tanpa missing values pada saat awal load, namun beberapa kolom (Glucose, BloodPressure, SkinThickness, Insulin, BMI) perlu diperhatikan karena nilai 0 yang tidak logis telah diimputasi dengan median.
- Semua fitur yang digunakan bersifat numerik sehingga tidak memerlukan encoding variabel kategori.
- Beberapa fitur numerik (terutama BMI dan Insulin) mengandung outlier yang bisa berdampak pada hasil modeling.
- Data sudah siap digunakan untuk tahap pra-pemrosesan lanjutan dan pembangunan model prediksi diabetes.

Data Preparation

Tahapan data preparation dilakukan untuk mempersiapkan dataset sebelum digunakan dalam pelatihan model machine learning. Berikut langkah-langkahnya:

1. **Penanganan Nilai Tidak Logis dan Imputasi**
   Beberapa fitur seperti Glucose, BloodPressure, SkinThickness, Insulin, dan BMI memiliki kemungkinan nilai 0 yang secara medis tidak valid. Nilai 0 pada kolom-kolom tersebut dikonversi menjadi NaN, kemudian diimputasi dengan nilai median kolom terkait untuk menjaga integritas data dan mencegah bias pada model.

---

2. **Pemisahan Fitur dan Target**
   Dataset dipisahkan menjadi dua bagian utama:
   - **Fitur (X):** Semua kolom kecuali target (Outcome).
   - **Target (y):** Kolom Outcome sebagai label target untuk klasifikasi (0 = tidak diabetes, 1 = diabetes).
     Pemisahan ini memastikan input data dan label target dapat digunakan secara terpisah dalam proses pelatihan dan evaluasi model.

---

3. **Normalisasi Data**
   Normalisasi dilakukan menggunakan **StandardScaler** sehingga semua fitur numerik memiliki skala yang seragam (mean = 0, std = 1). Langkah ini penting untuk memastikan setiap fitur memberikan kontribusi yang seimbang pada proses pembelajaran model machine learning.

---

4. **Pembagian Data Latih dan Data Uji**
   Dataset dibagi menjadi dua bagian:
   - **Data latih (80%)** digunakan untuk melatih model.

- **Data uji (20%)** digunakan untuk mengevaluasi performa model. Pembagian dilakukan secara acak (random_state=42) untuk menjaga distribusi kelas target tetap representatif pada kedua bagian data.

Kesimpulan Data Preparation

Setelah melalui tahapan ini, dataset menjadi:

- Bebas dari nilai tidak logis pada fitur utama (sudah diimputasi dengan median).
- Seluruh fitur numerik sudah ternormalisasi.
- Sudah terpisah menjadi data latih dan data uji dengan distribusi target yang proporsional.

Dataset yang telah dipersiapkan dengan langkah ini memastikan kualitas data yang optimal untuk proses modeling dan evaluasi model machine learning.

Modeling

Pada tahap ini, dilakukan pemodelan data menggunakan algoritma Random Forest untuk memprediksi risiko diabetes berdasarkan data kesehatan pasien. Model dilatih menggunakan data latih yang telah melalui proses pra-pemrosesan, kemudian digunakan untuk melakukan prediksi pada data uji. Evaluasi performa model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Selain itu, dilakukan analisis feature importance untuk mengidentifikasi fitur-fitur yang paling berpengaruh dalam menentukan prediksi diabetes pada dataset ini.

Random Forest

Random Forest adalah algoritma ensemble yang membangun banyak decision tree untuk menghasilkan prediksi yang lebih akurat dan stabil. Model ini bekerja dengan menggabungkan hasil dari beberapa pohon keputusan yang dilatih secara acak, sehingga mampu mengurangi risiko overfitting dan dapat menangkap interaksi kompleks antar fitur.

Parameter:

- random_state=42 untuk memastikan hasil yang konsisten pada setiap eksekusi model.

Hasil Evaluasi:

Berdasarkan data uji, performa model Random Forest adalah sebagai berikut:

- **Akurasi:** 0.74
- **Precision:**
    - Kelas 0 (Tidak Diabetes): 0.80
    - Kelas 1 (Diabetes): 0.63
- **Recall:**
    - Kelas 0: 0.79

    - Kelas 1: 0.65
- **F1 Score:**
    - Kelas 0: 0.80
    - Kelas 1: 0.64

Macro avg F1-score: 0.72
Weighted avg F1-score: 0.74

Confusion Matrix:

- True Negatives (0→0): 78
- False Positives (0→1): 21
- False Negatives (1→0): 19
- True Positives (1→1): 36

Kelebihan:

- Mampu menangani data dengan banyak fitur dan pola non-linear.
- Tahan terhadap overfitting karena menggunakan banyak pohon keputusan.
- Dapat mengukur tingkat kepentingan setiap fitur terhadap hasil prediksi (feature importance).

Kekurangan:

- Waktu pelatihan dan prediksi lebih lama dibanding model sederhana seperti Logistic Regression.
- Hasil prediksi lebih sulit diinterpretasikan secara langsung (kurang transparan dibanding model linear).

Cara Kerja:

- Model membangun banyak decision tree dari subset data dan fitur yang dipilih secara acak.
- Setiap pohon menghasilkan prediksi, dan hasil akhir ditentukan berdasarkan voting mayoritas dari seluruh pohon.
- Model juga dapat menghitung seberapa besar kontribusi setiap fitur terhadap prediksi akhir.

Analisis Feature Importance:

Dari hasil analisis, fitur paling penting dalam prediksi diabetes adalah Glucose, diikuti oleh BMI, Age, dan DiabetesPedigreeFunction. Fitur-fitur lain seperti Insulin, BloodPressure, SkinThickness, dan Pregnancies memiliki kontribusi yang lebih kecil.

Evaluation

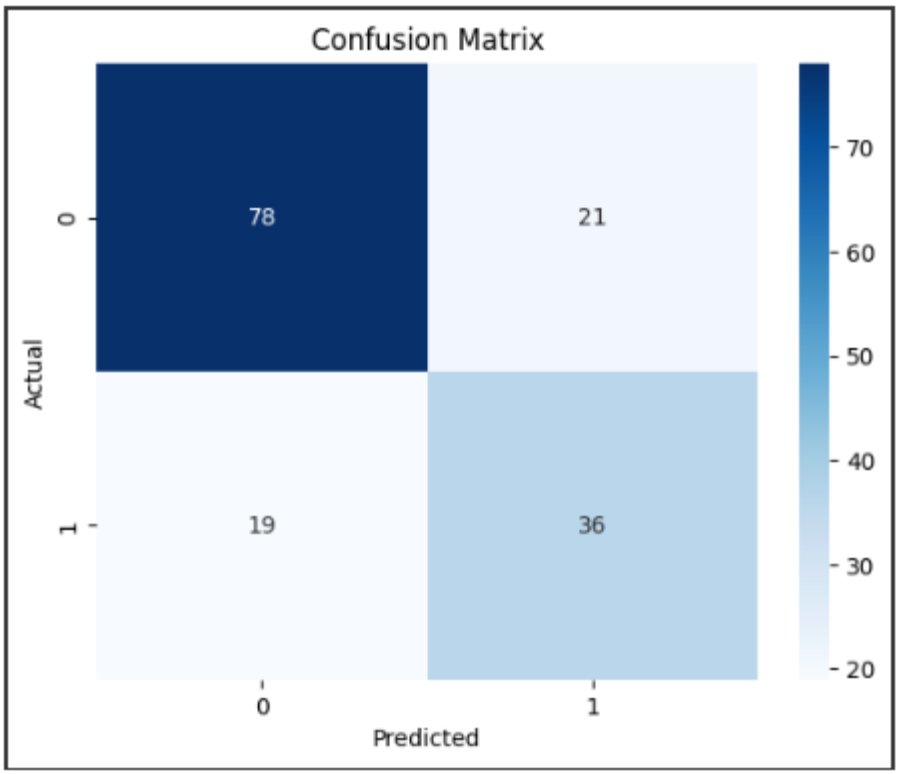
Pada tahap evaluasi, performa model Random Forest diukur menggunakan berbagai metrik yaitu Akurasi, Precision, Recall, dan F1 Score. Selain itu, confusion matrix digunakan untuk melihat detail prediksi benar dan salah pada masing-masing kelas.

Metrik Evaluasi

Metrik	Kelas 0 (Tidak Diabetes)	Kelas 1 (Diabetes)	Rata-rata (weighted)
Precision	0.80	0.63	0.74
Recall	0.79	0.65	0.74
F1 Score	0.80	0.64	0.74
Support	99	55	154
Akurasi			0.74

Analisis Confusion Matrix

Confusion matrix memberikan informasi tentang prediksi benar (True Positives dan True Negatives) serta prediksi salah (False Positives dan False Negatives). Berikut adalah confusion matrix dari masing-masing model:



- True Negatives (TN): 78
- False Positives (FP): 21
- False Negatives (FN): 19
- True Positives (TP): 36

Analisis:

- Model Random Forest menunjukkan performa baik dalam memprediksi diabetes, dengan akurasi sebesar **74%** pada data uji.
- Precision dan recall untuk kelas diabetes (1) berada pada angka 0.63 dan 0.65, menandakan model cukup baik dalam mendeteksi kasus positif meski masih terdapat beberapa kasus yang terlewat (FN).
- Nilai F1 Score yang seimbang antara kedua kelas menunjukkan model cukup stabil.

Berdasarkan analisis feature importance, fitur Glucose, BMI, dan Age merupakan faktor yang paling berpengaruh terhadap prediksi risiko diabetes.

Kesimpulan:
Model Random Forest yang digunakan memiliki keseimbangan performa antara akurasi, precision, recall, dan F1 score. Dengan performa yang stabil serta interpretasi fitur yang jelas, model ini layak dijadikan baseline untuk pengembangan prediksi risiko diabetes pada dataset ini.