

Implementasi Random Forest untuk Prediksi Risiko Diabetes Berbasis Gejala Klinis Menggunakan Analisis Feature Importance

^{1*}Frisqy Maulana, ²Risa Helilintar, ³Rony Hery Irawan

^{1,2,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: ¹fmkiki11@gmail.com, ²risa.helilintar@gmail.com, ³rony@unpkediri.ac.id

Penulis Korespondens : Frisqy Maulana

Abstrak— Diabetes melitus merupakan penyakit kronis yang memerlukan deteksi dini untuk mengurangi risiko komplikasi. Penelitian ini bertujuan mengimplementasikan algoritma *Random Forest* untuk prediksi risiko diabetes berbasis gejala klinis serta menganalisis faktor-faktor yang berpengaruh terhadap hasil prediksi menggunakan metode Feature Importance. Dataset yang digunakan adalah Early Stage Diabetes Risk Prediction Dataset yang terdiri dari 520 data. Tahapan penelitian meliputi preprocessing data, pembagian data dengan rasio 80:20, pelatihan model *Random Forest*, evaluasi model, dan analisis Feature Importance. Hasil penelitian menunjukkan bahwa model memperoleh akurasi 99,04%, precision 0,99, recall 0,99, F1-score 0,99, dan ROC-AUC 1,0000. Analisis Feature Importance menunjukkan bahwa Polyuria, Polydipsia, Gender, dan Age merupakan faktor yang paling berpengaruh terhadap prediksi risiko diabetes. Hasil penelitian menunjukkan bahwa *Random Forest* mampu memberikan performa prediksi yang sangat baik sekaligus mengidentifikasi faktor risiko dominan diabetes.

Kata Kunci— diabetes, feature importance, machine learning, random forest, prediksi risiko

Abstract— *Diabetes mellitus is a chronic disease that requires early detection to reduce the risk of complications. This study aims to implement the Random Forest algorithm for diabetes risk prediction based on clinical symptoms and to analyze influential factors using the Feature Importance method. The dataset used was the Early Stage Diabetes Risk Prediction Dataset consisting of 520 records. The research stages included data preprocessing, data splitting with an 80:20 ratio, model training, model evaluation, and Feature Importance analysis. The model achieved an accuracy of 99.04%, precision of 0.99, recall of 0.99, F1-score of 0.99, and ROC-AUC of 1.0000. Feature Importance analysis revealed that Polyuria, Polydipsia, Gender, and Age were the most influential factors in predicting diabetes risk. The results indicate that Random Forest provides excellent predictive performance while identifying dominant diabetes risk factors.*

Keywords— diabetes, feature importance, machine learning, random forest, risk prediction

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Diabetes melitus merupakan salah satu penyakit kronis yang menjadi masalah kesehatan global karena jumlah penderitanya terus meningkat setiap tahun. Penyakit ini ditandai dengan tingginya kadar glukosa dalam darah akibat gangguan produksi insulin atau ketidakmampuan tubuh dalam menggunakan insulin secara efektif. Apabila tidak ditangani dengan baik, diabetes dapat menyebabkan berbagai komplikasi serius, seperti penyakit kardiovaskular, gangguan ginjal,

kerusakan saraf, dan gangguan penglihatan. Oleh karena itu, deteksi dini terhadap risiko diabetes menjadi langkah penting untuk membantu proses pencegahan dan penanganan penyakit secara lebih efektif serta mengurangi risiko komplikasi yang dapat terjadi [1]. Upaya pemanfaatan teknologi untuk mendukung deteksi dini diabetes telah dilakukan melalui pengembangan sistem diagnosis berbasis metode K-NN yang mampu membantu proses identifikasi penyakit berdasarkan gejala yang dialami pasien [2].

Perkembangan teknologi informasi telah mendorong pemanfaatan *machine learning* dalam bidang kesehatan, khususnya untuk membantu proses prediksi dan deteksi dini penyakit. Melalui kemampuan dalam mempelajari pola dari data historis, *machine learning* dapat digunakan untuk mengidentifikasi individu yang berisiko mengalami diabetes berdasarkan faktor-faktor kesehatan yang dimiliki. Penelitian yang dilakukan oleh Deberneh dan Kim menunjukkan bahwa penerapan algoritma *machine learning* mampu memberikan performa yang baik dalam memprediksi diabetes tipe 2 sehingga berpotensi membantu proses pengambilan keputusan dalam bidang kesehatan [3].

Salah satu algoritma yang banyak digunakan dalam prediksi diabetes adalah *Random Forest*. Algoritma ini merupakan metode *ensemble learning* yang bekerja dengan membangun sejumlah *decision tree* dan menghasilkan keputusan berdasarkan suara mayoritas dari seluruh pohon yang terbentuk. Penelitian yang dilakukan oleh Noviyanti dan Alamsyah menunjukkan bahwa algoritma *Random Forest* mampu memberikan performa yang baik dalam proses deteksi dini diabetes karena memiliki kemampuan klasifikasi yang tinggi serta mampu menangani data dengan banyak atribut secara efektif [4].

Penelitian lain yang dilakukan oleh Anggreini et al. berfokus pada peningkatan performa prediksi diabetes menggunakan algoritma *Random Forest* dengan pendekatan *hyperparameter tuning*. Hasil penelitian menunjukkan bahwa proses optimasi parameter mampu meningkatkan kinerja model dalam mengklasifikasikan risiko diabetes. Temuan tersebut menunjukkan bahwa *Random Forest* merupakan algoritma yang memiliki performa yang baik dan masih dapat ditingkatkan melalui penyesuaian parameter yang tepat sehingga menghasilkan tingkat akurasi yang lebih optimal dalam proses prediksi [5].

Selain peningkatan performa model, penelitian mengenai seleksi fitur juga menjadi perhatian dalam pengembangan sistem prediksi diabetes. Rahman et al. mengkaji penggunaan berbagai algoritma *feature selection* dan *boosting-based Machine learning classifiers* untuk meningkatkan kualitas prediksi diabetes. Hasil penelitian menunjukkan bahwa pemilihan fitur yang relevan dapat memberikan kontribusi terhadap peningkatan performa model sekaligus membantu mengidentifikasi atribut yang memiliki pengaruh besar dalam proses klasifikasi. Temuan tersebut menunjukkan bahwa analisis terhadap kontribusi masing-masing fitur menjadi aspek penting yang tidak hanya berfokus pada akurasi model, tetapi juga pada pemahaman faktor-faktor yang memengaruhi hasil prediksi [6].

Penelitian yang dilakukan oleh Kaliappan et al. menganalisis berbagai strategi klasifikasi dan seleksi fitur dalam prediksi diabetes untuk mengetahui kombinasi metode yang mampu menghasilkan performa terbaik. Hasil penelitian menunjukkan bahwa pemilihan fitur yang tepat berpengaruh terhadap kinerja model prediksi dan dapat membantu mengurangi kompleksitas data

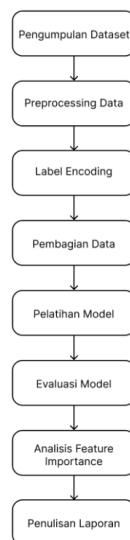
tanpa menghilangkan informasi penting. Meskipun demikian, sebagian besar penelitian masih berfokus pada peningkatan akurasi model dan penggunaan metode seleksi fitur, sementara analisis mengenai tingkat kontribusi masing-masing faktor risiko terhadap hasil prediksi masih relatif terbatas [7].

Berdasarkan penelitian-penelitian sebelumnya, dapat diketahui bahwa algoritma *machine learning*, khususnya *Random Forest*, memiliki performa yang baik dalam prediksi diabetes. Namun, sebagian besar penelitian lebih menitikberatkan pada peningkatan akurasi model melalui optimasi parameter maupun seleksi fitur. Analisis terhadap kontribusi setiap faktor risiko dalam proses prediksi masih belum banyak dibahas secara mendalam. Oleh karena itu, penelitian ini mengimplementasikan algoritma *Random Forest* untuk prediksi risiko diabetes berbasis gejala klinis serta melakukan analisis *Feature Importance* guna mengidentifikasi faktor-faktor yang paling berpengaruh terhadap hasil prediksi. Penelitian ini menggunakan *Early Stage Diabetes Risk Prediction* Dataset yang diperoleh dari *Kaggle* dengan tujuan mengidentifikasi faktor-faktor yang paling berpengaruh terhadap prediksi risiko diabetes serta mengevaluasi performa model yang dihasilkan.

II. METODE

Penelitian ini menggunakan pendekatan *machine learning* untuk mengimplementasikan algoritma *Random Forest* dalam prediksi risiko diabetes berbasis gejala klinis serta menganalisis faktor-faktor yang berpengaruh terhadap hasil prediksi menggunakan metode *Feature Importance*. Dataset yang digunakan adalah *Early Stage Diabetes Risk Prediction* Dataset yang diperoleh dari *Kaggle* dan terdiri dari 520 data dengan 16 atribut faktor risiko serta 1 atribut kelas sebagai target prediksi. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, pembagian data latih dan data uji, pembangunan model *Random Forest*, evaluasi performa model, serta analisis *Feature Importance*.

A. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1. Penelitian diawali dengan pengumpulan dataset yang diperoleh dari *Kaggle*. Selanjutnya dilakukan *preprocessing* data dan transformasi atribut kategorikal menggunakan teknik *Label Encoding*. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20. Data latih digunakan untuk membangun model *Random Forest*, sedangkan data uji digunakan untuk mengevaluasi performa model. Tahap terakhir dilakukan analisis *Feature Importance* untuk mengidentifikasi faktor-faktor risiko yang paling berpengaruh terhadap prediksi diabetes.

B. Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah *Early Stage Diabetes Risk Prediction Dataset* yang diperoleh dari *Kaggle*. Dataset terdiri dari 520 data dengan 16 atribut faktor risiko dan 1 atribut kelas (*class*) sebagai target prediksi. Atribut yang digunakan meliputi *Age*, *Gender*, *Polyuria*, *Polydipsia*, *Sudden Weight Loss*, *Weakness*, *Polyphagia*, *Genital Thrush*, *Visual Blurring*, *Itching*, *Irritability*, *Delayed Healing*, *Partial Paresis*, *Muscle Stiffness*, *Alopecia*, dan *Obesity*. Dataset ini digunakan untuk membangun model prediksi risiko diabetes menggunakan algoritma *Random Forest*.

C. Preprocessing Data

Tahap *preprocessing* dilakukan untuk memastikan kualitas data sebelum digunakan dalam proses pelatihan model. Pada penelitian ini dilakukan pemeriksaan terhadap struktur data, tipe data, dan keberadaan *missing value*. Hasil pemeriksaan menunjukkan bahwa dataset tidak memiliki nilai yang hilang sehingga data dapat langsung digunakan pada tahap berikutnya. Selanjutnya atribut kategorikal dikonversi ke dalam bentuk numerik menggunakan teknik *Label Encoding* agar dapat diproses oleh algoritma *Random Forest*.

D. Pembagian Data

Dataset yang telah melalui tahap *preprocessing* dibagi menjadi data latih (*training data*) dan data uji (*testing data*) menggunakan metode *train-test split*. Pembagian dilakukan dengan rasio 80:20 sehingga diperoleh 416 data latih dan 104 data uji. Data latih digunakan untuk membangun model *Random Forest*, sedangkan data uji digunakan untuk mengevaluasi performa model yang dihasilkan.

E. Pelatihan Model

Random Forest merupakan salah satu algoritma *ensemble learning* yang bekerja dengan membangun sejumlah *decision tree* dan menggabungkan hasil prediksi dari setiap pohon keputusan untuk menghasilkan keputusan akhir. Algoritma ini dipilih karena memiliki kemampuan klasifikasi yang baik, mampu menangani data dengan jumlah atribut yang banyak, serta memiliki ketahanan terhadap *overfitting* [8]. Pada penelitian ini, model *Random Forest* dilatih menggunakan data latih yang telah melalui tahap *preprocessing* dan *Label Encoding*. Proses pelatihan dilakukan untuk mempelajari pola hubungan antara faktor-faktor risiko dengan kelas diabetes sehingga model dapat digunakan untuk melakukan prediksi terhadap data baru [5].

Performa model dievaluasi menggunakan metrik akurasi yang dihitung menggunakan Persamaan (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Keterangan :

- Accuracy : tingkat ketepatan model dalam melakukan klasifikasi.

- TP (True Positive) : jumlah data positif yang diprediksi positif dengan benar.
- TN (True Negative) : jumlah data negatif yang diprediksi negatif dengan benar.
- FP (False Positive) : jumlah data negatif yang diprediksi positif.
- FN (False Negative) : jumlah data positif yang diprediksi negatif.

F. Implementasi dan Evaluasi Model

Model *Random Forest* yang telah dibangun kemudian diimplementasikan untuk melakukan prediksi risiko diabetes. Performa model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC. Selanjutnya dilakukan analisis *Feature Importance* untuk mengetahui tingkat kontribusi masing-masing atribut terhadap hasil prediksi diabetes.

Nilai *Feature Importance* dihitung menggunakan Persamaan (2).

$$FI_j = \frac{1}{T} \sum_{t=1}^T I_{jt} \quad (2)$$

Keterangan :

- FI_j : nilai *Feature Importance* untuk fitur ke- j .
- T : jumlah *decision tree* yang digunakan dalam model *random forest*.
- I_{jt} : kontribusi fitur ke- j pada pohon keputusan ke- t .

III. HASIL DAN PEMBAHASAN

A. Hasil Pembagian Data

Dataset yang digunakan dalam penelitian ini terdiri dari 520 data dengan 16 atribut faktor risiko dan 1 atribut kelas sebagai target prediksi. Setelah melalui tahap *preprocessing*, dataset dibagi menjadi data latih dan data uji menggunakan metode *train-test split* dengan rasio 80:20. Hasil pembagian menghasilkan 416 data latih dan 104 data uji. Data latih digunakan dalam proses pelatihan model *Random Forest*, sedangkan data uji digunakan untuk mengevaluasi performa model yang dihasilkan

Tabel 1. Pembagian Dataset

<i>Jenis Data</i>	<i>Jumlah</i>
Data Latih	416
Data Uji	104
Total	520

Berdasarkan Tabel 1, jumlah data latih yang digunakan lebih banyak dibandingkan data uji untuk memastikan model memperoleh informasi yang cukup selama proses pelatihan. Pembagian data dengan rasio 80:20 dipilih karena mampu memberikan keseimbangan antara proses pelatihan dan evaluasi model.

B. Hasil Evaluasi Model

Hasil pengujian model *Random Forest* menggunakan data uji menunjukkan performa klasifikasi yang sangat baik. Berdasarkan hasil evaluasi, model memperoleh nilai akurasi sebesar 99,04%. Selain itu, nilai *precision*, *recall*, dan *F1-score* masing-masing sebesar 0,99. Hasil tersebut menunjukkan bahwa model mampu mengklasifikasikan data risiko diabetes

dengan tingkat ketepatan yang sangat tinggi. Implementasi *Random Forest* dilakukan dengan menerapkan model yang telah dilatih pada data uji untuk menghasilkan prediksi risiko diabetes berdasarkan gejala klinis yang dimiliki setiap data pasien.

Model *Random Forest* yang telah dilatih menggunakan 416 data latih kemudian diimplementasikan untuk melakukan prediksi terhadap 104 data uji. Hasil prediksi yang diperoleh digunakan untuk mengukur performa model menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC.

Tabel 2. Hasil Evaluasi Model

<i>Metrik</i>	<i>Nilai</i>
Accuracy	99,04%
Precision	0,99
Recall	0,99
F1-Score	0,99
ROC-AUC	1,0000

Tingginya nilai akurasi menunjukkan bahwa algoritma *Random Forest* mampu mempelajari pola hubungan antara faktor-faktor risiko dan kelas diabetes dengan baik. Nilai *precision* sebesar 0,99 menunjukkan bahwa hampir seluruh data yang diprediksi berisiko diabetes merupakan data yang benar. Sementara itu, nilai *recall* sebesar 0,99 menunjukkan kemampuan model dalam mendeteksi kasus diabetes secara efektif. Nilai *F1-score* sebesar 0,99 mengindikasikan keseimbangan yang baik antara *precision* dan *recall*.

Hasil evaluasi menunjukkan bahwa model *Random Forest* memiliki performa klasifikasi yang sangat baik dalam memprediksi risiko diabetes. Hasil tersebut menunjukkan bahwa algoritma *Random Forest* efektif digunakan untuk prediksi risiko diabetes berbasis gejala klinis pada dataset yang digunakan. Perbedaan performa tersebut dapat dipengaruhi oleh karakteristik dataset, proses *preprocessing*, serta pemilihan parameter model yang digunakan. Dengan akurasi sebesar 99,04%, model yang dihasilkan mampu memberikan performa yang sangat baik dalam mengklasifikasikan risiko diabetes [4], [5].

C. Analisis Confusion Matrix

Confusion Matrix digunakan untuk mengetahui kemampuan model dalam mengklasifikasikan data berdasarkan kelas sebenarnya dan hasil prediksi yang dihasilkan. Berdasarkan hasil pengujian, model menghasilkan *Confusion Matrix* seperti yang ditunjukkan pada Tabel 3.

Tabel 3. Confusion Matrix

<i>Aktual/Prediksi</i>	<i>Negatif</i>	<i>Positif</i>
Negatif	40	0
Positif	1	63

Berdasarkan Tabel 3, model *Random Forest* berhasil mengklasifikasikan 40 data negatif dan 63 data positif dengan benar. Selain itu, tidak terdapat data negatif yang salah diklasifikasikan sebagai positif (*False Positive*). Namun, terdapat 1 data positif yang diprediksi sebagai negatif (*False Negative*). Hasil tersebut menunjukkan bahwa model

memiliki tingkat kesalahan yang sangat rendah dalam melakukan klasifikasi risiko diabetes sehingga mampu menghasilkan nilai akurasi yang tinggi.

D. Analisis *Feature Importance*

Analisis *Feature Importance* dilakukan untuk mengetahui tingkat kontribusi masing-masing faktor risiko terhadap hasil prediksi yang dihasilkan oleh model *Random Forest*. Semakin besar nilai *Feature Importance*, semakin besar pengaruh atribut tersebut dalam proses klasifikasi risiko diabetes. [6]

Tabel 4. Hasil Analisis *Feature Importance*

<i>Faktor Risiko</i>	<i>Importance</i>
Polyuria	0.204932
Polydipsia	0.180052
Gender	0.117120
Age	0.111317
Sudden Weight Loss	0.056890
Partial Paresis	0.046034
Alopecia	0.043785
Irritability	0.036365
Visual Blurring	0.030040
Delayed Healing	0.029854
Polyphagia	0.029453
Itching	0.028196
Genital Thrush	0.026895
Muscle Stiffness	0.021371
Weakness	0.020247
Obesity	0.017450

Berdasarkan hasil analisis *Feature Importance*, faktor Polyuria memiliki nilai kontribusi tertinggi sebesar 0,2049. Faktor berikutnya adalah Polydipsia sebesar 0,1801, diikuti oleh Gender sebesar 0,1171 dan Age sebesar 0,1113. Hasil tersebut menunjukkan bahwa gejala sering buang air kecil dan sering merasa haus merupakan faktor yang paling berpengaruh dalam proses prediksi risiko diabetes. Temuan ini sejalan dengan karakteristik klinis diabetes yang umumnya ditandai oleh peningkatan frekuensi buang air kecil dan rasa haus yang berlebihan [9].

Selain itu, faktor Gender dan Age juga menunjukkan kontribusi yang cukup tinggi dalam proses klasifikasi. Hal ini menunjukkan bahwa karakteristik demografis turut memengaruhi kemungkinan seseorang mengalami diabetes. Sementara itu, atribut seperti Obesity, Weakness, dan Muscle Stiffness memiliki nilai *Feature Importance* yang relatif rendah sehingga kontribusinya terhadap keputusan model tidak sebesar faktor-faktor utama lainnya. Hasil ini menunjukkan bahwa model mampu mengidentifikasi faktor risiko yang paling dominan [10] dan memberikan informasi yang dapat digunakan sebagai dasar dalam upaya deteksi dini diabetes. Temuan ini sejalan dengan penelitian Hasan dkk. yang menunjukkan bahwa analisis

Feature Importance dapat meningkatkan transparansi model prediksi diabetes dan membantu memahami faktor-faktor yang berpengaruh terhadap hasil prediksi [11].

E. Implementasi Sistem

Model *Random Forest* yang telah dilatih kemudian diimplementasikan ke dalam aplikasi berbasis *web* menggunakan *Streamlit*. Sistem memungkinkan pengguna memasukkan data gejala klinis untuk memperoleh hasil prediksi risiko diabetes secara otomatis berdasarkan model yang telah dibangun.

Prediksi Risiko Diabetes

Masukkan data pasien untuk melakukan prediksi

[Kembali ke Beranda](#)

⚠ Hasil prediksi ini bukan diagnosis medis final. Silakan konsultasikan dengan tenaga kesehatan profesional.

Umur (Age)
30
Masukkan umur Anda saat ini dalam satuan tahun.

Jenis Kelamin (Gender)
Laki-laki
Pilih jenis kelamin sesuai dengan identitas biologis Anda.

Sering Buang Air Kecil (Polyuria)
Iya
Buang air kecil lebih sering dari biasanya, terutama pada malam hari, meskipun tidak sedang banyak minum.

Sering Haus (Polydipsia)
Iya
Merasa haus berlebihan dan ingin minum terus-menerus meskipun sudah cukup minum.

Penurunan Berat Badan Drastis (Sudden Weight Loss)
Iya
Mengalami penurunan berat badan yang cukup signifikan tanpa melakukan diet atau olahraga khusus.

Tubuh Lemah / Mudah Lelah (Weakness)
Iya
Sering merasa lemas, kurang bertenaga, atau mudah lelah saat melakukan aktivitas sehari-hari.

Nafsu Makan Berlebih (Polyphagia)
Iya
Nafsu makan meningkat atau merasa lapar lebih sering dibandingkan biasanya.

Infeksi Jamur Kelamin (Genital Thrush)
Iya
Mengalami infeksi jamur pada area kelamin yang ditandai dengan gatal, kemerahan, atau rasa tidak nyaman.

Penglihatan Kabur (Visual Blurring)
Iya
Penglihatan menjadi kabur atau tidak sejelas biasanya saat melihat objek.

Gatal-gatal (Itching)
Iya
Mengalami rasa gatal pada kulit yang sering muncul tanpa penyebab yang jelas.

Mudah Marah / Sensitif (Irritability)
Iya
Lebih mudah marah, sensitif, atau merasa emosional dibandingkan biasanya.

Luka Sulit Sembuh (Delayed Healing)
Iya
Luka kecil atau goresan membutuhkan waktu lebih lama untuk sembuh dibandingkan biasanya.

Kelemahan Otot Sebagian (Partial Paresis)
Iya
Mengalami kelemahan pada sebagian anggota tubuh sehingga gerakan terasa tidak sekuat biasanya.

Kaku Otot (Muscle Stiffness)
Iya
Otot terasa kaku atau sulit digerakkan, terutama setelah beristirahat atau bangun tidur.

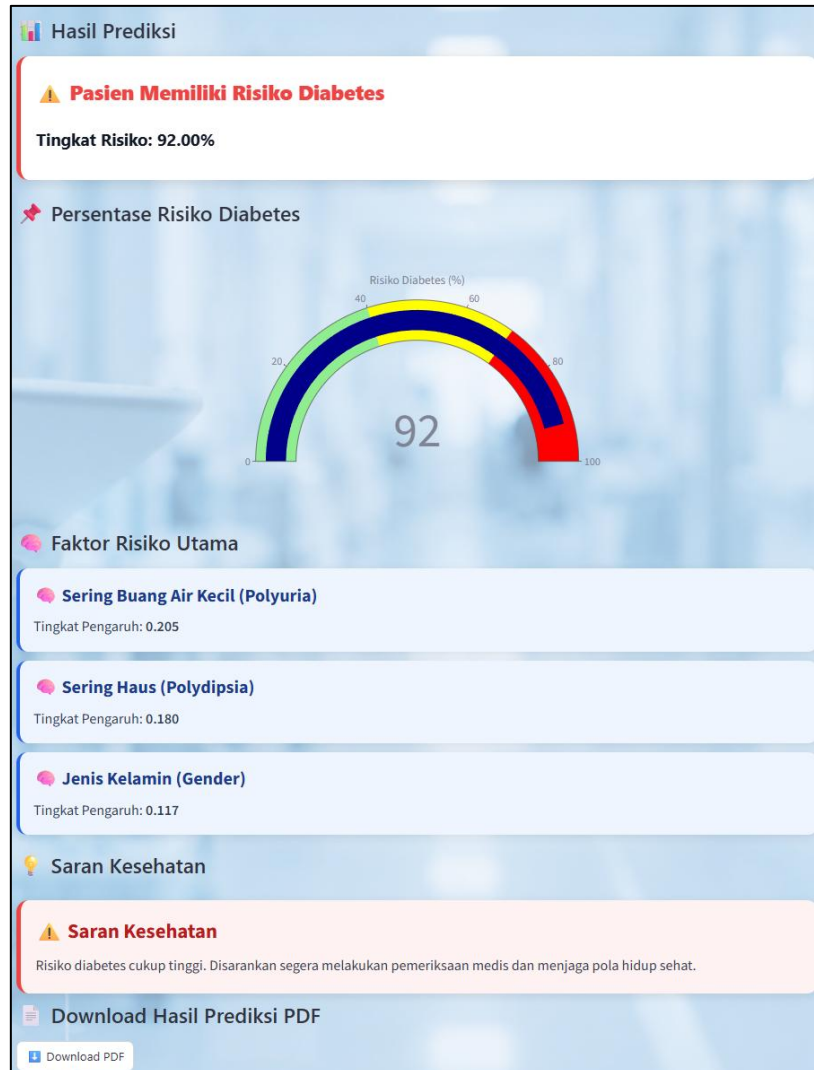
Kerontokan Rambut (Alopecia)
Iya
Mengalami kerontokan rambut yang lebih banyak dari biasanya hingga menyebabkan penipisan rambut.

Obesitas (Obesity)
Iya
Memiliki berat badan berlebih yang dapat meningkatkan risiko berbagai penyakit, termasuk diabetes.

[Prediksi Risiko Diabetes](#)

Gambar 2. Halaman Input Data Prediksi Risiko Diabetes

Gambar 2 menunjukkan halaman input yang digunakan pengguna untuk memasukkan data gejala klinis yang diperlukan dalam proses prediksi risiko diabetes. Data yang dimasukkan kemudian diproses oleh model *Random Forest* yang telah dilatih sebelumnya.



Gambar 3. Hasil Prediksi Risiko Diabetes

Gambar 3 menunjukkan hasil prediksi yang dihasilkan oleh sistem setelah pengguna memasukkan seluruh data gejala klinis. Sistem menampilkan hasil klasifikasi risiko diabetes beserta tingkat probabilitas prediksi berdasarkan model *Random Forest* yang digunakan.

IV. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma *Random Forest* untuk prediksi risiko diabetes berbasis gejala klinis menggunakan *Early Stage Diabetes Risk Prediction Dataset*. Hasil pengujian menunjukkan bahwa model mampu menghasilkan akurasi sebesar 99,04%, *precision* 0,99, *recall* 0,99, *F1-score* 0,99, dan ROC-AUC 1,0000. Analisis *Feature Importance* menunjukkan bahwa *Polyuria*, *Polydipsia*, *Gender*, dan *Age* merupakan faktor yang memberikan kontribusi terbesar terhadap hasil prediksi. Hasil penelitian menunjukkan bahwa implementasi *Random Forest* mampu menghasilkan performa prediksi yang sangat baik sekaligus membantu mengidentifikasi faktor risiko dominan sebagai dasar pendukung deteksi dini diabetes.

DAFTAR PUSTAKA

- [1] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, pp. 9–11, 2021, doi: 10.3390/ijerph18063317.
- [2] R. A. Ramadhani and R. K. Niswatin, "Sistem Diagnosa Diabetes Menggunakan Metode K-NN," *Jurnal Sains dan Informatika*, vol. 4, no. 2, pp. 98–104, 2018, doi: 10.34128/jsi.v4i2.121.
- [3] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.
- [4] C. N. Noviyanti and A. Alamsyah, "Early Detection of Diabetes Using Random Forest Algorithm," *Journal of Information System Exploration and Research*, vol. 2, no. 1, pp. 41–48, 2024, doi: 10.52465/joiser.v2i1.245.
- [5] N. L. Anggreini, A. Yuliana, D. S. Ramdan, and W. Al-Dayyeni, "Improving Diabetes Prediction Performance Using Random Forest Classifier with Hyperparameter Tuning," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 1847–1860, 2025, doi: 10.52436/1.jutif.2025.6.4.4755.
- [6] F. Rahman, S. Hossain, J. J. Tiang, and A. Al Nahid, "Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers," *Diagnostics*, vol. 15, no. 20, pp. 1–24, 2025, doi: 10.3390/diagnostics15202622.
- [7] J. Kaliappan *et al.*, "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1421751.
- [8] M. Rizky, R. Kurniawan, and D. Lestari, "Deteksi Dini Diabetes Mellitus Menggunakan Algoritma Random Forest pada Data Klinis," *Journal of Computer Science and Information Technology*, vol. 2, no. 1, pp. 1–8, 2026, doi: 10.70716/jocsit.v2i1.410.
- [9] S. Majyambere, T. Lindgren, C. Twizere, and I. Ntakirutimana, "Early Type 2 diabetes risk prediction using explainable machine learning in a two-stage approach.," *Front. Digit. Health*, vol. 8, no. March, p. 1743619, 2026, doi: 10.3389/fdgth.2026.1743619.
- [10] D. N. Jawza, M. I. Mazdadi, A. Farmadi, T. H. Saragih, D. Kartini, and V. Abdullayev, "Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature Selection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 295–306, 2025, doi: 10.35882/jeeemi.v7i2.626.
- [11] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights," *Information (Switzerland)*, vol. 16, no. 1, 2025, doi: 10.3390/info16010007.