Jurnal Teknik dan Science
JTS Vol 4 No. 3 October 2025 | ISSN: 2828-7002 (Print), ISSN: 2828-6871 (online), Page 39-46

# Penggunaan Kecerdasan Buatan untuk Menganalisis Faktor Risiko Diabetes dengan menggunakan Random Forest Classifier

**Tri Sulistyorini[1], Nelly Sofi[2], Dwi Widiastuti[3], Viliananda Tripita Claur[4]**
[1,2,4]Informatika, Universitas Gunadarma
[3]Sistem Informasi, Universitas Gunadarma

**Article History**
Received : 15-08-2025
Revised : 20-08-2025
Accepted : 19-10-2025
Published : 27-10-2025

**Corresponding author*:**
tri_s@staff.gunadarma.ac.id

**Cite This Article:** Tri Sulistyorini, Nelly Sofi, Dwi Widiastuti, & Viliananda Tripita Claur. (2025). Penggunaan Kecerdasan Buatan untuk Menganalisis Faktor Risiko Diabetes dengan menggunakan Random Forest Classifier. Jurnal Teknik Dan Science, 4(3), 39–46.

**DOI:**
https://doi.org/10.56127/jts.v4i3.2453

**Abstract:** Diabetes is a non-communicable disease that deserves attention and poses a significant public health challenge. Although not a contagious disease, preventive measures and early detection of diabetes risk are crucial. This study used machine learning-based artificial intelligence to identify diabetes risk factors. The model was created using the Random Forest Classifier (RFC) algorithm, which has 16 variables as parameters. The model was built using the Python programming language, with data collection spanning from 2015 to 2018. The research included needs analysis, data collection, data preprocessing, model training, predictive model creation, system design, implementation, and testing. The final results showed that, with an accuracy of 89%, the model could be used effectively to predict diabetes risk. Furthermore, the model identified more pre-diabetes classes than other classes.

**Keywords:** Machine Learning, Random Forest, Diabetes Prediction, Python

## PENDAHULUAN

Diabetes salah satu penyakit yang tidak menular, namun menjadi tantangan besar bagi kesehatan masyarakat diberbagai negara. Di Indonesia, jumlah penderita diabetes sudah mencapai sekitar 10,3 juta orang, data didapat dari laporan International Diabetes Federation (IDF) pada tahun 202. Jumlah penderita ini sangat berkaitan erat dengan perubahan gaya hidup masyarakat modern, seperti pola makan yang buruk, kurangnya aktivitas fisik, serta kebiasaan merokok dan konsumsi alkohol yang tinggi. Kondisi ini memperlihatkan urgensi untuk melakukan upaya preventif dan deteksi dini terhadap risiko diabetes.

Perkembangan teknologi yang semakin meningkat, serta pemanfaatan kecerdasan buatan (AI) semakin luas, memungkinkan sistem untuk menganalisis data dalam jumlah besar untuk mengidentifikasi pola yang relevan dalam pengambilan keputusan.

Machine learning menjadi salah satu pendekatan yang potensial untuk memprediksi risiko diabetes berdasarkan data riwayat kesehatan individu.

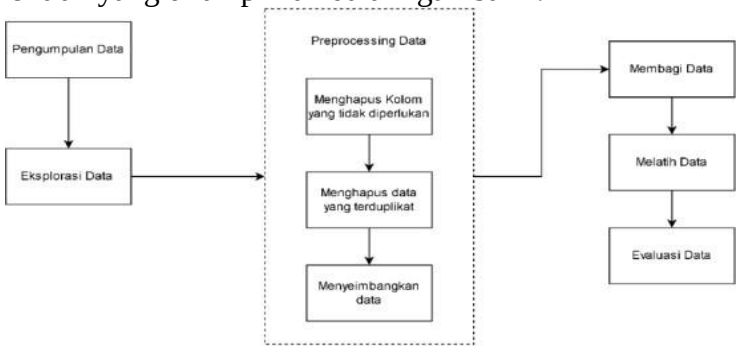
Teknologi ini dapat digunakan untuk membantu masyarakat mengetahui tingkat risiko secara cepat dan akurat.

Salah satu algoritma yang efektif dalam proses klasifikasi data medis adalah Random Forest Classifier, karena kemampuannya dalam menangani data kompleks dan menghasilkan prediksi yang stabil [6].

Penelitian ini bertujuan untuk membangun sebuah model prediksi risiko diabetes menggunakan algoritma Random Forest Classifier yang diimplementasikan melalui bahasa pemrograman Python.

## METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif melalui pemanfaatan data sekunder dari Kaggle yang memuat 16 variabel terkait kondisi kesehatan. Berikut ini adalah alur dari penelitian yang dilampirkan dalam gambar 1.




**Gambar 1.** Alur Pengerjaan Penelitian

Seluruh proses dilakukan secara bertahap, dimulai dari:

1. Analisis kebutuhan: Identifikasi perangkat keras dan lunak yang diperlukan selama proses penelitian.
2. Pengumpulan dan pra- pemrosesan data: Data pasien dikumpulkan dari internet, kemudian dibersihkan, dikategorikan, dan disesuaikan skalanya.
3. Pelatihan dan pembuatan model: Model dilatih menggunakan algoritma Random Forest Classifier untuk memprediksi status diabetes, pre-diabetes, atau normal.
4. Perancangan dan implementasi: Model yang telah dilatih diintegrasikan ke dalam aplikasi desktop.
5. Pengujian : Sistem diuji untuk memastikan akurasi prediksi.

## PEMBAHASAN

Penelitian ini mengembangkan sistem prediksi risiko diabetes dengan memanfaatkan algoritma Random Forest Classifier (RFC). Model ini dilatih dengan 16 variabel kesehatan dari dataset BRFSS yang diperoleh melalui Kaggle. Sistem ini dirancang agar dapat menerima input pengguna sesuai dengan format variabel model.

Pengumpulan Data

Pada tahap ini data yang digunakan merupakan data kesehatan yang didapat melalui Kaggle.

Data yang didapat memiliki 20 variabel bebas dan 1 variabel terikat, namun hanya 16 variabel yang digunakan untuk membuat model. Rentang waktu dari dataset ini adalah mulai dari tahun 2015 sampai dengan 2018. Berikut penjelasan dari masing – masing variable dalam table 1.

| Nama Variabel | Deskripsi | Nilai / Kategori |
|---|---|---|
| **Binary Diabetes** | Variabel target status diabetes | 0 = No Diabetes<br>1 = Prediabetes<br>2 = Diabetes |
| **HighBP** | Terdiagnosa tekanan darah tinggi | 0 = Tidak ada<br>1 = Ada darah tinggi |
| **HighChol** | Terdiagnosa kolesterol tinggi | 0 = Tidak ada<br>1 = Pernah terdiagnosa |
| **CholCheck** | Pemeriksaan kolesterol dalam 5 tahun terakhir | 0 = Tidak ada<br>1 = Ada |
| **Smoker** | Pernah merokok ≥ 100 batang seumur hidup | 0 = Tidak ada<br>1 = Ada |
| **Stroke** | Pernah terdiagnosa stroke | 0 = Tidak ada<br>1 = Ada |
| **HeartDiseaseorAttack** | Pernah terdiagnosa penyakit jantung koroner atau serangan jantung | 0 = Tidak ada<br>1 = Ada |
| **PhysActivity** | Aktivitas fisik dalam 30 hari terakhir (tidak termasuk pekerjaan) | 0 = Tidak ada<br>1 = Ada |
| **Fruits** | Mengonsumsi buah ≥ 1 kali per hari | 0 = Tidak ada<br>1 = Ada |
| **Veggies** | Mengonsumsi sayur ≥ 1 kali per hari | 0 = Tidak ada<br>1 = Ada |
| **GenHlth** | Status kesehatan secara umum | 1 = Sangat baik<br>2 = Baik<br>3 = Biasa saja<br>4 = Buruk<br>5 = Sangat buruk |
| **MentHlth** | Jumlah hari mengalami stres, depresi, atau masalah emosi dalam 30 hari terakhir | 0 – 30 hari |
| **PhysHlth** | Jumlah hari mengalami gangguan fisik, sakit, atau cedera dalam 30 hari terakhir | 0 – 30 hari |
| **DiffWalk** | Kesulitan berjalan atau menaiki tangga | 0 = Tidak ada<br>1 = Ada |
| **Sex** | Jenis kelamin responden | 0 = Wanita<br>1 = Pria |
| **Age** | Kelompok usia (13 kategori) | 1 = 18–24<br>2 = 25–29<br>…<br>13 = 60–64 |

| Nama Variabel | Deskripsi | Nilai / Kategori |
|---|---|---|
| **BMI** | Body Mass Index | Berat Badan / (Tinggi Badan)² |
| **HvyAlcoholConsump** | Konsumsi alkohol berat (Pria >14 kali/minggu, Wanita >7 kali/minggu) | 0 = Tidak ada<br>1 = Ada |

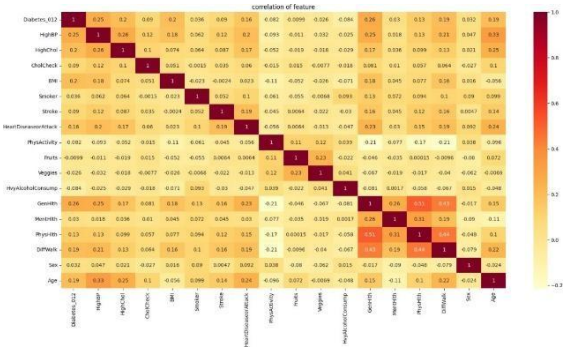
## Data Preprocessing

Langkah pertama yang dilakukan adalah pembersihan data, termasuk penghapusan variabel yang tidak relevan, pengecekan dan penghapusan data duplikat, serta penyesuaian tipe data pada beberapa kolom. Yang ditampilkan pada gambar 2.

```
df.drop(columns=['NoDocbcCost', 'AnyHealthcare', 'Income', 'Education'], inplace=True)
df.head(5)
```

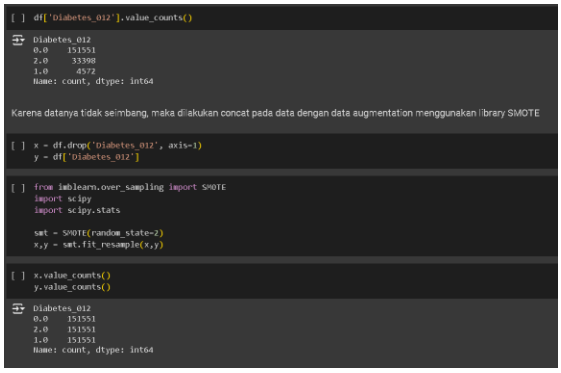
**Gambar 2.** Menghapus Kolom

Selanjutnya dilakukan analisis korelasi antar variabel dengan menggunakan heatmap. Dari hasil visualisasi, ditemukan bahwa variabel seperti Tekanan Darah Tinggi (highBP), BMI, Status Kesehatan, Usia, dan Penyakit Jantung memiliki korelasi positif yang signifikan terhadap risiko diabetes. Temuan ini sejalan dengan teori dalam dunia medis yang menyatakan bahwa faktor-faktor tersebut merupakan komponen penting dalam risiko diabetes tipe 2.




**Gambar 3**. Heatmap Korelasi

Jumlah data pada kelas non-diabetes jauh lebih banyak dibandingkan kelas diabetes atau pre-diabetes, Hal ini dikarenakan data awal yang memiliki distribusi kelas yang tidak seimbang, maka langkah selanjutnya yaitu menyeimbangkan data dengan menggunakan metode SMOTE (Synthetic Minority Oversampling Technique). Metode SMOTE menjadikan data minoritas diduplikasi secara sintetis untuk meningkatkan performa model dalam mengenali pola pada seluruh kelas. Penggunaan dan hasil dari SMOTE ini dapat dilihat pada gambar 4.

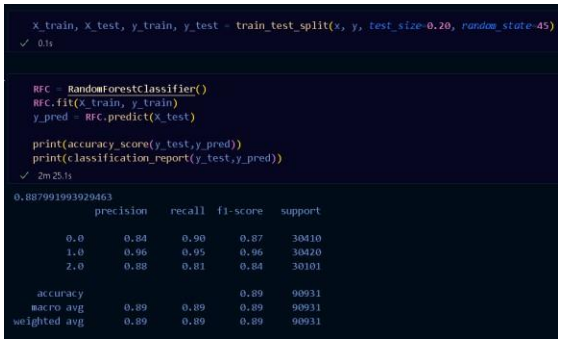



**Gambar 4.** Penyeimbangan data SMOTE

**Pembuatan dan Evaluasi Model**

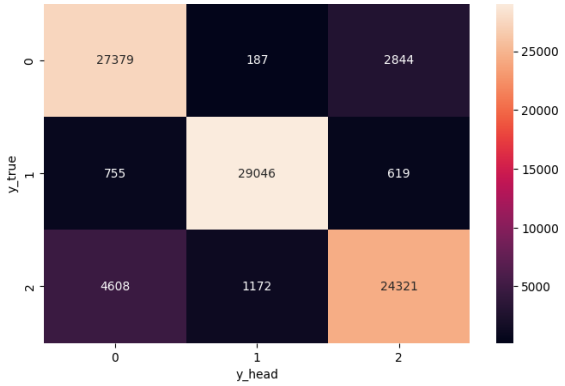
Model yang digunakan dalam penelitian ini adalah Random Forest Classifier (RFC), sebuah metode ensemble learning yang membentuk banyak pohon keputusan dan menggabungkan hasilnya untuk mendapatkan prediksi yang lebih stabil dan akurat. Model ini dipilih karena memiliki performa yang baik dalam klasifikasi multi- kelas dan relatif robust terhadap overfitting.

Data dibagi menjadi dua bagian, yaitu 20% untuk pengujian (testing) dan 80% untuk pelatihan (training). Proses pelatihan dilakukan dengan menyesuaikan parameter model secara default, dan hasil akurasi yang diperoleh mencapai 89%, yang menunjukkan bahwa model cukup andal dalam memprediksi risiko diabetes berdasarkan input variabel yang diberikan. Hal ini dapat dilihat pada gambar 5 yang dilampirkan hasil perhitungan dari classification report model.




**Gambar 5.** Classification Report RFC

Performa model dapat dievaluasi menggunakan confusion matrix. Hasil evaluasi menunjukkan bahwa model mampu memprediksi kelas pre-diabetes dengan lebih baik dibandingkan kelas lainnya, meskipun terdapat sedikit kesalahan klasifikasi antara kelas non-diabetes dan pre-diabetes. Hal ini cukup wajar mengingat keduanya sering memiliki gejala awal yang serupa secara klinis. Selain confusion matrix, akurasi, presisi, recall, dan f1-score juga dihitung untuk memberikan gambaran yang lebih komprehensif terhadap performa model yang telah dilampirkan di gambar 5. Confusion matrix pada gambar 6 digunakan melihat berapa nilai True dan Falsenya.




**Gambar 6.** Confusion Matrix

Berdasarkan confusion matrix pada gambar 6 maka hasil analisis model terhadap data testing adalah sebagai berikut:

1. True non-diabetes sebanyak 27379 data
2. True pre-diabetes sebanyak 29046 data
3. True diabetes sebanyak 24321 data
4. False non-diabetes sebanyak 187 dan 2844 data
5. False pre-diabetes sebanyak 755 dan 619 data
6. False diabetes sebanyak 4608 dan 1172 data

**Implementasi Prediksi Risiko Diabetes**

Setelah model berhasil dibangun dan diuji, tahap selanjutnya adalah implementasi model ke dalam bentuk desktop. Aplikasi ini terdiri atas beberapa menu, sbb :

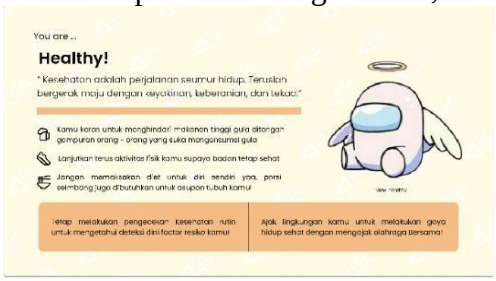
• Menu Utama : berisikan penjelasan singkat mengenai tujuan dari aplikasi. Rancangan terlihat pada gambar 7 berikut :


Gambar 7. Wireframe Landing Page

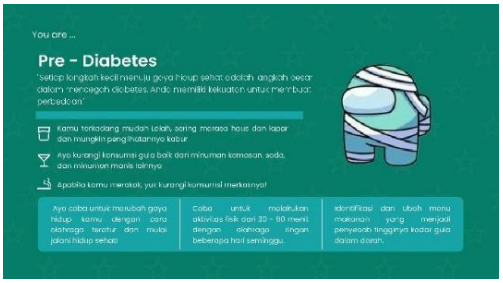
• Menu Test : berisikan sebuah form yang harus diisi pengguna untuk mendapatkan hasil prediksi, dengan beberapa isian seperti : usia, BMI, status merokok. Tampilan Form untuk test terlihat pada gambar 8 berikut :


**Gambar 8.** Rancangan Form

• Hasil Prediksi akan muncul secara otomatis jika pengguna mengklik submit di form test (gambar 8), yang akan menampilkan hasil klasifikasi dari model dalam bentuk gambar yang secara berturut – turut akan dilampirkan dalam gambar 9, 10 dan 11.




**Gambar 9.** Hasil Sehat




**Gambar 10**. Hasil Pre - Diabetes


**Gambar 11.** Hasil Diabetes

## KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa sistem prediksi risiko diabetes berbasis kecerdasan buatan menggunakan Random Forest Classifier didapat nilai akurasi sebesar 89% yang menunjukkan bahwa model memiliki kinerja yang cukup baik dalam memprediksi risiko diabetes pada data pasien.

Pengembangan lebih lanjut dapat dilakukan pada sistem ini seperti dari sisi fungsi prediksi, sehingga menjadi alat bantu sebagai bagian dari solusi digital preventif dalam dunia kesehatan

## DAFTAR PUSTAKA

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer, 21(3), 677–690

Gunay, Denis. Random Forest, https://medium.com/@denizgunay/ random-forest-af5bde5d7e1e, diakses pada tanggal 24 Juni 2025.

IBM. What is machine learning, https://www.ibm.com/topics/machi ne-learning diakses pada tanggal 25 Juni 2024.

Loke, A. 2023. "Diabetes", diakses pada tanggal 24 Juni 2025, https://www.who.int/news-room/fact-sheets/detail/diabetes.

MayoClinic Staff. 2023. "Prediabetes", https://www.mayoclinic.org/disease s-conditions/prediabetes/symptoms- causes/syc-20355278, diakses pada tanggal 24 Juni 2025.

Suryady, S., Soerowirdjo, B., & Sari, S. P. (2025). Impact of Combining RGB and Grayscale Images on Hotspot Detection in Solar Panels Using Inception Resnet V2 Architecture. Ingenierie des Systemes d'Information, 30(4), 1043.

Hasan, A., Ali K., Amani S. (2024). Random Forest Algorithm Overview, https://mesopotamian.press/journals/index.php/BJML, diakses 3 April 2025

Jakin V., Vimal S., Kaliappan M., Young L. (2021), AI based smart prediction of clinical disease using random forest classifier and Naive Bayes, diakses pada tanggal 12 April 2025.

Yohanes, Robertus. 2022. Diabetes Melitus, "Ibu dari Berbagai Penyakit", https://herminahospitals.com/id/articles/diabetes-melitus-ibu-dari-berbagai-penyakit.html, diakses pada tanggal 24 Juni 2025.