

BAB I

PENDAHULUAN

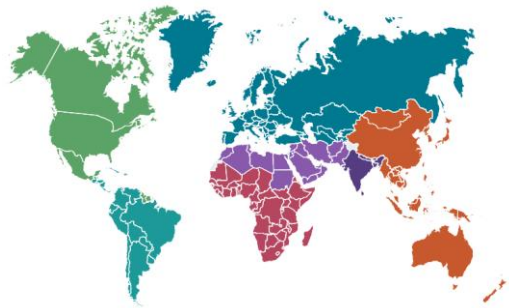
1.1 Latar Belakang

Diabetes merupakan penyakit kronis yang kini menjadi salah satu tantangan terbesar dalam dunia kesehatan global. Kondisi ini terjadi ketika tubuh tidak mampu memproduksi insulin secara memadai atau ketika insulin yang dihasilkan tidak dapat bekerja secara optimal dalam mengendalikan kadar gula darah. Ketidakmampuan tersebut menyebabkan penumpukan glukosa dalam aliran darah yang berpotensi memicu berbagai komplikasi serius, mulai dari gangguan jantung, kerusakan saraf, penurunan fungsi ginjal, hingga risiko kematian. Besarnya dampak yang ditimbulkan menjadikan diabetes sebagai prioritas dalam berbagai upaya pencegahan dan penanganan kesehatan hingga saat ini.

Prevalensi diabetes secara global menunjukkan tren peningkatan yang konsisten dari tahun ke tahun. Berdasarkan data yang dipublikasikan oleh *International Diabetes Federation* [1] dalam *Diabetes Atlas* edisi ke-11, jumlah penderita diabetes di seluruh dunia pada tahun 2024 telah mencapai 589 juta orang dalam rentang usia 20–79 tahun, dan angka ini diproyeksikan akan melonjak hingga 853 juta jiwa pada tahun 2050. Peningkatan tersebut erat kaitannya dengan perubahan pola hidup masyarakat modern, seperti minimnya aktivitas fisik, konsumsi makanan tinggi gula, kebiasaan merokok, serta kurangnya waktu tidur yang diidentifikasi sebagai faktor risiko utama. Gambaran tren peningkatan jumlah penderita diabetes secara global dapat dilihat pada Gambar 1.1. *Worldwide Data Diabetes*.

Summary

Map 1 Number of people with diabetes worldwide and per IDF Region, in 2024–2050 (20–79 years)



Gambar 1. 1 Worldwide Data Diabetes

Jumlah penderita diabetes terus mengalami peningkatan di berbagai wilayah di dunia, termasuk pada tingkat global dan regional. Data terkait jumlah penderita diabetes secara global dan di wilayah Asia Tenggara pada tahun 2024 serta proyeksi peningkatannya hingga tahun 2050 disajikan pada Tabel 1.1. Data Penderita Diabetes Global dan Asia Tenggara dan pada Gambar 1.2 *World Data Diabetes*.

Tabel 1. 1 Data Penderita Diabetes Global dan Asia Tenggara

Wilayah	Jumlah Penderita 2024	Proyeksi 2050	Kenaikan
Global	589 juta jiwa	853 juta jiwa	45%
Asia Tenggara	106,9 juta jiwa	184,5 juta jiwa	73%



Gambar 1. 2 World Data Diabetes

Indonesia turut menghadapi beban penyakit tidak menular yang signifikan, salah satunya diabetes dengan jumlah penderita yang tergolong tinggi. Pada regional Asia Tenggara, termasuk Indonesia IDF [1] memproyeksikan jumlah penderita diabetes akan melonjak sebesar 73% dari 106,9 juta pada 2024 menjadi 184,5 juta jiwa pada 2050 mengacu pada Gambar 1.3 *South-East Asia Data Diabetes*, salah satu kenaikan tertinggi di dunia. Kondisi ini semakin mengkhawatirkan mengingat 43%

penderita diabetes belum terdiagnosis, sehingga diperlukan upaya deteksi dini yang lebih efektif agar penanganan dapat dilakukan lebih cepat.



Gambar 1. 3 South-East Asia Data Diabetes

Kemajuan teknologi komputasi membuka peluang pemanfaatan *Machine Learning* sebagai pendekatan analitis dalam bidang kesehatan, termasuk untuk memprediksi risiko diabetes pada individu. Melalui pembelajaran pola dari kumpulan data yang tersedia, sistem berbasis *machine learning* mampu melakukan klasifikasi maupun prediksi terhadap kondisi kesehatan tertentu secara lebih akurat. Prediksi tersebut mempertimbangkan berbagai faktor risiko yang relevan, seperti usia, indeks massa tubuh (BMI), tingkat aktivitas fisik, dan sejumlah indikator gaya hidup lainnya.

Sejumlah penelitian sebelumnya telah mengeksplorasi penerapan algoritma *machine learning* dalam prediksi diabetes dengan berbagai pendekatan. Algoritma seperti *Decision Tree*, *Random Forest*, *K-Nearest Neighbor*, dan *Logistic Regression* telah menunjukkan kemampuan yang baik dalam melakukan klasifikasi risiko diabetes. Dalam penelitian ini, algoritma *Logistic Regression* dipilih karena memiliki tingkat interpretabilitas yang tinggi serta banyak digunakan sebagai model dasar (*baseline model*) pada permasalahan klasifikasi biner di bidang kesehatan. Sementara itu, *Random Forest* dipilih karena mampu menangani hubungan non-linear antar variabel, lebih robust terhadap *overfitting*, serta dapat memberikan informasi mengenai tingkat kepentingan setiap fitur (*feature importance*) yang digunakan dalam proses prediksi.

Selain pemilihan algoritma, optimasi model juga menjadi faktor penting dalam meningkatkan kualitas prediksi. Oleh karena itu, penelitian ini menerapkan *hyperparameter tuning* menggunakan metode *GridSearchCV* untuk memperoleh

kombinasi parameter terbaik pada masing-masing algoritma. Pendekatan ini dilakukan untuk memastikan bahwa model yang digunakan tidak hanya mengandalkan konfigurasi standar, tetapi telah melalui proses optimasi yang sistematis sehingga mampu menghasilkan performa yang lebih baik.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan membandingkan model klasifikasi Diabetes Mellitus Tipe 2 menggunakan algoritma *Logistic Regression* dan *Random Forest* dengan pendekatan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) pada dataset kesehatan yang tersedia secara publik. Perbandingan performa kedua algoritma dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Selain melakukan evaluasi terhadap model yang dihasilkan, penelitian ini juga mengimplementasikan model terbaik ke dalam aplikasi web berbasis *Streamlit* sebagai bentuk *deployment* sehingga hasil penelitian dapat dimanfaatkan secara lebih interaktif. Melalui penelitian ini diharapkan diperoleh model yang mampu membantu proses deteksi dini diabetes secara lebih efektif serta memberikan gambaran mengenai faktor-faktor yang paling berpengaruh terhadap risiko diabetes.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana proses pembangunan model klasifikasi diabetes tipe 2 menggunakan algoritma *Logistic Regression* dan *Random Forest* dengan pendekatan CRISP-DM?
2. Bagaimana pengaruh penerapan *hyperparameter tuning* menggunakan *GridSearchCV* terhadap performa algoritma *Logistic Regression* dan *Random Forest* dalam mengklasifikasikan diabetes tipe 2?
3. Algoritma manakah yang memberikan performa terbaik dalam mengklasifikasikan diabetes tipe 2 berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score*?

1.3 Batasan Masalah

1. *Dataset* yang digunakan dalam penelitian ini berupa dataset *Diabetes Mellitus* Tipe 2 yang diperoleh dari Kaggle dengan total sebanyak 100.000 data.

2. Algoritma yang dikaji dalam penelitian ini dibatasi hanya *Logistic Regression* dan *Random Forest*.
3. Evaluasi kinerja model dilakukan dengan mengacu pada metrik *accuracy*, *precision*, *recall*, dan *F1-score*.
4. Penelitian mencakup implementasi model terbaik ke dalam aplikasi web berbasis *Streamlit* untuk melakukan prediksi risiko diabetes.
5. *Process preprocessing* data mencakup penanganan *missing values* dan encoding data kategorikal, serta normalisasi fitur numerik menggunakan *StandardScaler*.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Penelitian ini memiliki tujuan sebagai berikut:

1. Membangun model klasifikasi diabetes tipe 2 menggunakan algoritma *Logistic Regression* dan *Random Forest* dengan pendekatan CRISP-DM.
2. Menganalisis pengaruh penerapan *hyperparameter tuning* menggunakan *GridSearchCV* terhadap performa algoritma *Logistic Regression* dan *Random Forest*.
3. Menentukan algoritma dengan kinerja terbaik berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*.

1.4.2 Manfaat Penelitian

Penelitian ini memiliki manfaat sebagai berikut:

1. Manfaat Teoritis

Melalui penelitian ini diharapkan dapat menambah referensi mengenai penerapan, perbandingan, dan optimasi algoritma *machine learning* menggunakan *hyperparameter tuning* dalam klasifikasi diabetes tipe 2.

2. Manfaat Praktis

Diharapkan penelitian ini dapat memberikan referensi bagi tenaga kesehatan maupun peneliti dalam menentukan algoritma *machine learning* yang lebih efektif untuk klasifikasi dan deteksi dini risiko diabetes

berdasarkan indikator klinis dan gaya hidup. Selain itu, implementasi model ke dalam aplikasi web berbasis *Streamlit* diharapkan dapat membantu pengguna dalam melakukan prediksi risiko diabetes secara interaktif.

1.5 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran menyeluruh mengenai isi dan alur pembahasan setiap bab secara teratur. Adapun susunannya adalah sebagai berikut:

BAB I PENDAHULUAN Bab ini menguraikan latar belakang penelitian, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA Bab ini memaparkan landasan teori yang mendasari penelitian, mencakup konsep diabetes, *machine learning*, klasifikasi, algoritma *Logistic Regression*, algoritma *Random Forest*, metode evaluasi model, serta kajian penelitian terdahulu yang relevan.

BAB III METODOLOGI PENELITIAN Bab ini mendeskripsikan jenis penelitian, sumber data, variabel penelitian, tahapan penelitian, proses praproses data, penerapan algoritma *Logistic Regression* dan *Random Forest*, proses *hyperparameter tuning* menggunakan *GridSearchCV*, teknik evaluasi model, serta tahapan *deployment* menggunakan *Streamlit*.

BAB IV HASIL DAN PEMBAHASAN Bab ini menyajikan hasil pengolahan data, hasil *hyperparameter tuning*, evaluasi model, analisis perbandingan performa algoritma *Logistic Regression* dan *Random Forest*, analisis *feature importance*, serta implementasi model ke dalam aplikasi web berbasis *Streamlit*.

BAB V PENUTUP Bab ini memuat kesimpulan yang ditarik berdasarkan hasil penelitian serta saran yang dapat dijadikan pertimbangan bagi pengembangan penelitian selanjutnya.