

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka dilakukan untuk mengidentifikasi perkembangan penelitian yang relevan serta menentukan posisi penelitian yang dilakukan. Perbandingan antara penelitian terdahulu dan penelitian ini ditinjau berdasarkan judul penelitian, metode yang digunakan, sistem yang dikembangkan, serta hasil penelitian. Ringkasan perbandingan tersebut disajikan pada Tabel 2.1.

Tabel 2. 1 Perbandingan Penelitian

Peneliti	Algoritma	Dataset	Hasil Penelitian	Perbedaan Penelitian
(Huwaidi et al., 2022)	<i>K-Nearest Neighbor</i>	Pima Indians Diabetes Dataset	Sistem deteksi diabetes berbasis web dengan <i>K-Nearest Neighbor</i>	Hanya satu algoritma dan penelitian ini membandingkan <i>Naive Bayes</i> dan <i>Decision Tree</i>
Fauzi (2025)	<i>Naive Bayes</i>	Dataset Kaggle (diabetes)	Aplikasi sistem pakar deteksi diabetes dengan <i>Naive Bayes</i>	Hanya satu algoritma, tanpa perbandingan performa
Nuryamin dan Risyda (2025)	<i>Decision Tree C4.5</i> dan <i>Naive Bayes</i>	Dataset Kaggle (diabetes)	Komparasi <i>Decision Tree C4.5</i> dan <i>Naive Bayes</i> dengan RapidMiner	Hanya evaluasi model pada penelitian ini mengimplementasikan ke sistem berbasis web
Permana & Patwari (2021)	<i>Naive Bayes</i> dan <i>Decision Tree</i>	Dataset Kaggle (diabetes)	Komparasi kedua algoritma untuk prediksi diabetes dengan RapidMiner	Belum diimplementasikan ke sistem berbasis web
Penelitian ini (2026)	<i>Naive Bayes</i> dan <i>Decision Tree</i>	Dataset Kaggle (diabetes)	Komparasi kedua algoritma dan implementasi model terbaik ke sistem berbasis web	Menggabungkan proses komparasi algoritma dengan implementasi sistem prediksi diabetes berbasis web

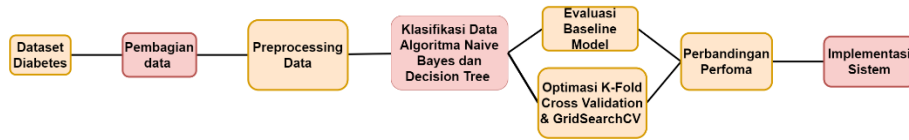
Berdasarkan Tabel 2.1, penelitian prediksi diabetes menggunakan metode *machine learning* telah banyak dilakukan dengan berbagai algoritma, seperti K-Nearest Neighbor, *Naive Bayes*, dan *Decision Tree*. Beberapa penelitian sebelumnya telah berhasil menerapkan algoritma klasifikasi untuk prediksi diabetes, namun masih terdapat keterbatasan pada aspek implementasi maupun perbandingan model.

Penelitian (Huwaiddi et al., 2022) dan (Fauzi, 2025) telah mengembangkan sistem prediksi diabetes berbasis web, tetapi masih menggunakan satu algoritma sehingga belum memberikan informasi mengenai perbandingan performa antar algoritma. Sementara itu, penelitian oleh (Nuryamin & Risyda, 2025) serta (Permana & Patwari, 2021) telah melakukan komparasi algoritma *Naive Bayes* dan *Decision Tree*, tetapi masih berfokus pada evaluasi performa model tanpa implementasi ke dalam sistem berbasis web.

Berdasarkan keterbatasan tersebut, penelitian ini dilakukan dengan menggabungkan proses komparasi algoritma *Naive Bayes* dan *Decision Tree* serta implementasi sistem berbasis web. Penelitian ini diharapkan dapat memberikan informasi mengenai algoritma dengan performa terbaik dalam prediksi diabetes serta menyediakan sistem yang dapat menampilkan hasil prediksi dan perbandingan performa algoritma secara lebih mudah.

2.2 Kerangka Pemikiran

Di dalam sebuah penelitian, kerangka pemikiran menjadi dasar dalam menentukan arah masalah yang akan dikaji dan menggambarkan alur pelaksanaan penelitian secara sistematis. Kerangka pemikiran dalam penelitian ini disusun untuk memberikan gambaran logis mengenai proses penelitian, mulai dari penggunaan dataset hingga implementasi sistem (Aulia, 2024). Flowchart Kerangka pemikiran penelitian ini ditunjukkan pada Gambar 2.1



Gambar 2. 1 Flowchart Kerangka Pemikiran penelitian

Penelitian ini menggunakan *Pima Indians Diabetes Dataset* sebagai sumber data yang kemudian dilakukan pembagian data dengan rasio 80:20, dan *preprocessing* untuk mempersiapkan data sebelum proses klasifikasi. Selanjutnya, algoritma *Naive Bayes* dan *Decision Tree* diterapkan untuk membangun model prediksi diabetes.

Kedua algoritma tersebut dievaluasi melalui model *baseline* dan dilakukan optimasi menggunakan *Stratified K-Fold Cross Validation* serta *GridSearchCV* untuk memperoleh konfigurasi model terbaik. Performa kedua algoritma kemudian dibandingkan berdasarkan metrik evaluasi yang digunakan. Model dengan performa terbaik selanjutnya diterapkan pada sistem berbasis web untuk melakukan prediksi status diabetes berdasarkan data pasien.

2.3 Landasan Teori

Landasan Teori ini membahas teori yang mendukung penelitian tentang prediksi penyakit diabetes menggunakan algoritma *Naive Bayes* dan *Decision Tree* berbasis web. Pembahasan meliputi diabetes, *data mining*, *Machine learning*, *Supervised Learning* klasifikasi, *preprocessing* data, *Naive Bayes*, *Decision Tree*, *Confusion Matrix*, *Cross Validation*, dan *Hyperparameter Tuning*.

a. Diabetes

Diabetes merupakan penyakit kronis yang ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan metabolisme tubuh. Berdasarkan data *World Health Organization (WHO)*, jumlah penderita diabetes mengalami peningkatan dari 108 juta orang pada tahun 1980 menjadi 422 juta orang pada tahun 2014 (*World Health Organization*, sebagaimana dikutip dalam Cendani & Wibowo, 2022).

Berdasarkan penyebabnya, diabetes dibedakan menjadi tiga jenis, yaitu diabetes tipe 1, diabetes tipe 2, dan diabetes gestasional. Diabetes tipe 1 terjadi karena tubuh tidak mampu menghasilkan insulin, sedangkan diabetes tipe 2 disebabkan oleh resistensi insulin atau gangguan penggunaan insulin. Sementara itu, diabetes gestasional terjadi selama masa kehamilan.

Beberapa faktor yang dapat meningkatkan risiko diabetes meliputi usia, riwayat keluarga, hipertensi, kurangnya aktivitas fisik, dan tingginya indeks massa tubuh (*Body Mass Index/BMI*). Diagnosis diabetes umumnya didasarkan pada beberapa indikator klinis yang memiliki nilai ambang batas tertentu.

Pada indeks massa tubuh (BMI), seseorang masih dikategorikan normal apabila nilainya tidak melebihi $24,9 \text{ kg/m}^2$. Apabila BMI berada pada rentang $25\text{--}29,9 \text{ kg/m}^2$, maka dikategorikan *overweight*, dan apabila mencapai 30 kg/m^2 atau lebih, maka dikategorikan obesitas, yang merupakan salah satu faktor risiko utama diabetes tipe 2 (World Health Organization, 2025).

Pada kadar glukosa darah puasa, batas maksimal yang masih dikategorikan normal adalah $70\text{--}100 \text{ mg/dL}$, diukur setelah puasa selama minimal delapan jam. Apabila kadar glukosa darah puasa mencapai atau melebihi 126 mg/dL , kondisi tersebut sudah mengarah pada diagnosis diabetes (American Diabetes Association, 2023).

Untuk tekanan darah, batas atas yang masih dianggap sehat adalah di bawah $120/80 \text{ mmHg}$. Tekanan darah dikategorikan mulai meningkat apabila mencapai $120\text{--}129/<80 \text{ mmHg}$, dan dikategorikan hipertensi apabila mencapai $\geq 130/80 \text{ mmHg}$, di mana kondisi ini banyak dialami oleh penderita diabetes melitus (American Diabetes Association, 2017, sebagaimana dikutip dalam (Nuraeni et al., 2026)).

Adapun untuk kadar insulin, nilai insulin puasa yang masih dikategorikan normal berada pada rentang $2\text{--}25 \text{ } \mu\text{IU/mL}$. Nilai yang melebihi batas atas tersebut dapat mengindikasikan resistensi insulin, yang berkaitan erat dengan risiko diabetes tipe 2.

Dalam penelitian ini, data kesehatan pasien meliputi kadar glukosa, tekanan darah, insulin, dan BMI digunakan sebagai dasar untuk membangun model prediksi diabetes menggunakan metode *machine learning*.

b. *Data Mining*

Data mining merupakan proses untuk menemukan pola atau informasi dari kumpulan data dengan memanfaatkan teknik statistik, basis data, dan *machine learning*. Dalam bidang kesehatan, *data mining* dapat digunakan untuk menganalisis data pasien serta menemukan pola yang mendukung proses pengambilan Keputusan (Irwandi et al., 2025).

Salah satu teknik dalam *data mining* adalah klasifikasi, yaitu proses pengelompokan data berdasarkan karakteristik tertentu ke dalam kelas yang telah ditentukan.

c. *Machine learning*

Machine learning merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem mempelajari pola dari data untuk menghasilkan prediksi atau keputusan tanpa pemrograman secara langsung (Ningsih et al., 2022).

Pada penelitian ini, *machine learning* digunakan untuk membangun model prediksi diabetes dengan menerapkan algoritma *Naive Bayes* dan *Decision Tree* sebagai metode klasifikasi.

d. *Supervised Learning*

Supervised learning merupakan metode dalam *machine learning* yang menggunakan data berlabel untuk melatih model. Data tersebut terdiri dari variabel *input* berupa atribut dan *output* berupa kelas target yang akan diprediksi (Dinata & Hasdyna, 2025).

Penelitian ini menggunakan *dataset* yang memiliki label status diabetes berupa positif dan negatif. Oleh karena itu, pendekatan *supervised learning* digunakan agar model dapat mempelajari hubungan antara atribut pasien dengan hasil klasifikasi diabetes.

e. Klasifikasi

Klasifikasi merupakan salah satu metode dalam *supervised learning* yang digunakan untuk mengelompokkan data ke dalam kelas tertentu berdasarkan pola pada atribut yang dimiliki (Srirahayu & Pribadie, 2023). Dalam penelitian ini, metode klasifikasi digunakan untuk membangun model prediksi diabetes berdasarkan pola dari atribut kesehatan pasien.

f. *Preprocessing Data*

Preprocessing data merupakan tahap awal dalam *data mining* yang bertujuan untuk membersihkan dan mempersiapkan data sebelum dilakukan proses analisis (Daniswara & Nuryana, 2023). Tahapan preprocessing diperlukan karena data mentah dapat memiliki nilai yang tidak sesuai atau perbedaan skala antar atribut yang dapat memengaruhi proses pembelajaran model.

Pada penelitian ini, proses *preprocessing* meliputi perubahan nilai 0 yang tidak valid menjadi *missing value*, pengisian nilai yang hilang menggunakan median, pengecekan kembali kondisi data, serta proses standardisasi fitur numerik menggunakan *StandardScaler*.

g. *Naive Bayes*

Naive Bayes adalah algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes untuk memperkirakan kelas suatu data. Algoritma ini mengasumsikan bahwa setiap atribut bersifat independen terhadap kelas yang diprediksi, sehingga perhitungannya menjadi lebih sederhana dan ringan secara komputasi. Meski sederhana, Naive Bayes terbukti mampu memberikan hasil yang baik pada berbagai permasalahan klasifikasi (Kawani, 2019). Teorema Bayes dirumuskan sebagai berikut:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (1)$$

Dimana menyatakan H menyatakan hipotesis atau kelas tertentu, X adalah data atau fitur yang diamati, $P(H|X)$ merupakan probabilitas posterior, $P(H)$ adalah probabilitas prior, $P(X|H)$ adalah likelihood, dan $P(X)$ adalah evidence.

Proses klasifikasi dilakukan dengan menghitung nilai probabilitas data untuk setiap kelas, kemudian data tersebut dikategorikan ke dalam kelas dengan nilai probabilitas terbesar. Pada penelitian ini, Naive Bayes digunakan untuk memprediksi status diabetes berdasarkan atribut kesehatan pasien yang telah melalui tahap *preprocessing*. (Ridwan, 2020).

h. *Decision Tree*

Decision Tree adalah algoritma *supervised learning* untuk klasifikasi dan regresi, yang merepresentasikan proses pengambilan keputusan dalam bentuk pohon: node sebagai atribut, cabang sebagai hasil pengujian atribut, dan leaf node sebagai hasil akhir klasifikasi. (Nurani et al., 2023; Permana & Patwari, 2021).

Pemilihan atribut pada setiap node dilakukan berdasarkan kemampuan atribut tersebut dalam membagi data menjadi kelas yang lebih homogen, menggunakan perhitungan *entropy* dan *information gain* untuk menentukan atribut terbaik sebagai pemisah data.

$$Entropy(S) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (2)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (3)$$

Pada penelitian ini, *Decision Tree* digunakan sebagai algoritma pembandingan terhadap *Naive Bayes* dalam melakukan prediksi diabetes berdasarkan atribut kesehatan pasien.

i. *Confusion Matrix*

Confusion matrix merupakan metode evaluasi yang digunakan untuk membandingkan hasil prediksi model dengan kondisi sebenarnya. Evaluasi ini terdiri dari empat komponen, yaitu *True Positive* (TP), *True Negative* (TN),

False Positive (FP), dan *False Negative* (FN) sebagaimana disajikan pada Tabel 2.2

Tabel 2. 2 *Confusion Matrix*

Aktual / Prediksi	Prediksi Positif Diabetes	Prediksi Negatif Diabetes
Aktual Positif	<i>True Positive</i> (TP)	<i>False Negative</i> (FN)
Aktual Negatif	<i>False Positive</i> (FP)	<i>True Negative</i> (TN)

Berdasarkan Tabel 2.2 keempat komponen tersebut menjadi dasar perhitungan metrik evaluasi, seperti akurasi, sensitivitas, dan spesifisitas. Pada penelitian ini, *confusion matrix* digunakan untuk mengevaluasi serta membandingkan performa algoritma *Naive Bayes* dan *Decision Tree* dalam melakukan prediksi diabetes (Valerian et al., 2025).

j. *Cross Validation*

Cross validation merupakan metode evaluasi model dengan membagi data menjadi beberapa bagian (*fold*) untuk digunakan secara bergantian sebagai data pelatihan dan pengujian. Metode ini bertujuan memperoleh hasil evaluasi yang lebih stabil.

Penelitian ini menggunakan *K-Fold Cross Validation* dengan nilai $k = 5$. Data dibagi menjadi lima bagian dan proses pelatihan serta pengujian dilakukan secara bergantian pada setiap *fold*. Selain itu, penggunaan *Stratified K-Fold* dilakukan karena dataset memiliki dua kelas, yaitu diabetes dan non-diabetes, sehingga distribusi kelas perlu dipertahankan pada setiap pembagian data (Wijiyanto et al., 2024).

k. *Hyperparameter Tuning*

Hyperparameter tuning merupakan proses pencarian kombinasi parameter terbaik pada model untuk meningkatkan performa dan kestabilan hasil prediksi. Pada penelitian ini, metode yang digunakan adalah *GridSearchCV*, yaitu teknik pencarian kombinasi parameter terbaik berdasarkan rentang parameter yang telah ditentukan (Hidayat et al., 2026).