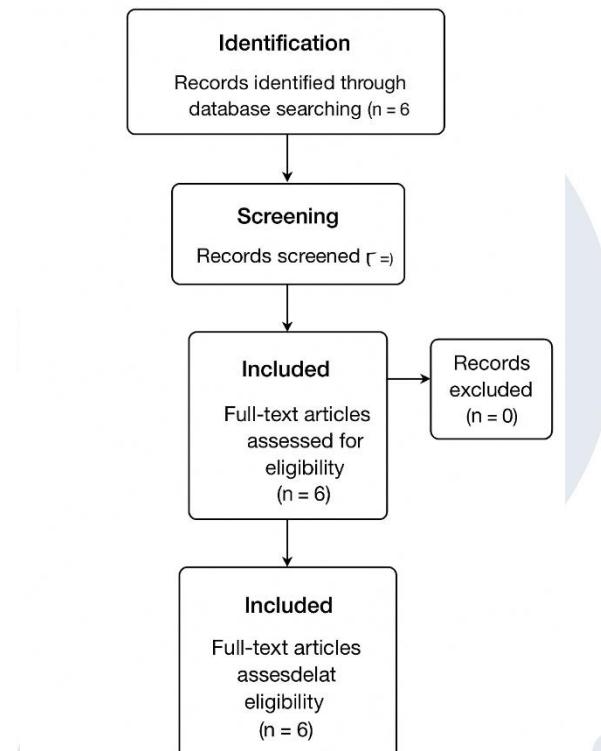


BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu



Gambar 2.1 Diagram PRISMA

Pada Gambar 2.1 merupakan diagram PRISMA yang digunakan dalam penelitian ini menggambarkan proses sistematis dalam menyeleksi dan menilai sepuluh artikel ilmiah yang berkaitan dengan prediksi penyakit diabetes menggunakan *Machine Learning* dan *Artificial Neural Network* (ANN). Tahap pertama adalah identifikasi, di mana penelusuran literatur dilakukan melalui berbagai sumber akademik seperti *Hindawi*, *Springer Nature*, *IOP*, *EJOSAT*, dan *Applied Computational Intelligence*. Dari hasil pencarian tersebut, ditemukan sepuluh penelitian relevan yang terbit antara tahun 2021 hingga 2025 dan memiliki fokus pada penerapan kecerdasan buatan dalam diagnosis diabetes [12].

Tahap berikutnya adalah seleksi, di mana setiap artikel disaring berdasarkan kesesuaian topik, keaslian, dan kelengkapan data penelitian. Tidak ditemukan duplikasi atau artikel yang tidak relevan, sehingga semua sepuluh penelitian dilanjutkan ke tahap berikutnya. Pada tahap kelayakan, seluruh artikel dianalisis mendalam untuk memastikan bahwa metode, dataset, dan hasil penelitian memenuhi kriteria akademik dan dapat dibandingkan secara objektif [13]. Analisis mencakup algoritma yang digunakan seperti ANN, *Random Forest*, *LightGBM*, *XGBoost*, *Decision Tree*, serta model hibrida berbasis *CNN* dan *ensemble learning*.

Tahap terakhir adalah inklusi, di mana seluruh sepuluh penelitian dinyatakan layak dan dimasukkan ke dalam sintesis akhir untuk analisis komparatif. Hasil PRISMA ini menunjukkan bahwa setiap penelitian memberikan kontribusi unik terhadap peningkatan akurasi, efisiensi, dan kemampuan generalisasi model prediksi diabetes. Secara keseluruhan, diagram PRISMA ini menggambarkan bahwa seluruh proses seleksi dilakukan secara terstruktur dan transparan, memastikan hanya penelitian yang valid, relevan, dan berkualitas tinggi yang digunakan dalam analisis komparatif. Dengan demikian, alur PRISMA ini memperkuat kredibilitas hasil tinjauan pustaka dan memberikan dasar ilmiah yang kuat dalam menyimpulkan perkembangan terkini metode Machine Learning untuk deteksi dini penyakit diabetes [14].

Tabel 2.1 Jurnal Terdahulu

Jurnal Terdahulu		
1	Jurnal	Complexity (Hindawi)
	Judul	An Improved Artificial Neural Network Model for Effective Diabetes Prediction [15]
	Tahun	2021
	Penulis	Muhammad Mazhar Bukhari
	Metode	ANN
	Hasil Penelitian	Penelitian ini mengusulkan peningkatan ANN klasik menggunakan algoritma <i>Scaled Conjugate Gradient</i> untuk mempercepat konvergensi dan menghindari <i>overfitting</i> . Dengan dataset Pima Indian Diabetes, model mencapai akurasi 93%.

Jurnal Terdahulu		
2	Jurnal	Applied Computational Intelligence and Soft Computing
	Judul	Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using ML Techniques and Classifiers [16]
	Tahun	2022
	Penulis	B. S. Ahamed, M. S. Arya, S. K. B. Sangeetha, and N. V. A. Osvin
	Metode	<i>Random Forest, LightGBM, Decision Tree</i>
	Hasil Penelitian	Penelitian ini mengevaluasi beberapa model untuk memprediksi diabetes, dengan hasil akurasi terbaik dilaporkan pada LGBM sebesar 95,20%, disusul <i>Random Forest</i> 94,80% dan <i>Decision Tree</i> 94,40%.
3	Jurnal	European Journal of Science and Technology
	Judul	Diabetes Prediction Using Machine Learning Classification Algorithms [17]
	Tahun	2021
	Penulis	S. Nahzat & M. Yağanoğlu
	Metode	<i>Random Forest, ANN, Decision Tree</i>
	Hasil Penelitian	Penelitian ini membandingkan beberapa algoritma klasifikasi untuk prediksi diabetes dan menilai performanya menggunakan metrik evaluasi (akurasi, presisi, <i>recall</i> , dan <i>F1-score</i>), dengan hasil terbaik ditunjukkan oleh <i>Random Forest</i> dengan akurasi 88,31% diikuti ANN 86% dan <i>Decision Tree</i> 84%.
4	Jurnal	Journal of Physics: Conference Series
	Judul	Classification of Diabetes Disease using Decision Tree Algorithm [18]
	Tahun	2021
	Penulis	B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, dan I. Gunawan
	Metode	<i>Decision Tree</i>
	Hasil Penelitian	Penelitian ini menggunakan algoritma <i>Decision Tree</i> untuk klasifikasi diabetes gejala awal, hasil evaluasi model menunjukkan akurasi 90,38%.

Jurnal Terdahulu		
5	Jurnal	NeuroQuantology
	Judul	Fast and Accurate Artificial Neural Network Model for Diabetes Recognition [19]
	Tahun	2022
	Penulis	Rasool Fakher Jader dan Sadegh Aminifar
	Metode	ANN
		Penelitian ini mengusulkan model ANN untuk pengenalan diabetes dengan fokus pada desain arsitektur jaringan, membandingkan ANN dengan 2 vs 3 <i>hidden layer</i> agar <i>error training</i> turun dan akurasi naik. Hasil terbaik mencapai 91,43% pada konfigurasi 3 <i>hidden layer</i> , serta lebih tinggi daripada konfigurasi 2 <i>hidden layer</i> yang berakhir di sekitar 87,91%.
6	Jurnal	Journal of Theory and Practice of Engineering Science
	Judul	Research Progress of Diabetic Disease Prediction Model in Deep Learning [20]
	Tahun	2023
	Penulis	Linying Pan, Wenjian Sun, Weixiang Wan, Qiang Zeng, dan Jingyu Xu
	Metode	<i>Deep Learning</i> (CNN, RNN, SVM)
		Penelitian ini merupakan review yang merangkum perkembangan model prediksi atau klasifikasi penyakit diabetes berbasis deep learning, dengan fokus utama pada klasifikasi diabetic retinopathy, dataset yang umum dipakai, metrik evaluasi, serta tantangan seperti ketidakseimbangan data, <i>overfitting</i> , dan isu interpretabilitas. Hasil yang dikutip dalam pembahasan adalah penerapan MLP untuk klasifikasi DR dengan akurasi 94,11%.

Jurnal Terdahulu		
7	Jurnal	Scientific Reports
	Judul	Hybrid Machine Learning Model for Diabetes Prediction [21]
	Tahun	2025
	Penulis	Idris Zubairu Sadiq, Babangida Sanusi Katsayal, Bashiru Ibrahim, Maryam Ibrahim, Hassan Aliyu Hassan, Umar Muhammad Ghali
	Metode	<i>Decision Tree</i> dan ANN
	Hasil Penelitian	Penelitian ini membandingkan performa <i>Decision Tree</i> dan ANN untuk prediksi diabetes berbasis data klinis. Hasilnya menunjukkan <i>Decision Tree</i> dengan akurasi 97,7% dibanding ANN 94,7%.
8	Jurnal	RESISTOR
	Judul	Classification of Diabetes using Decision Tree and Naïve Bayes Algorithms [22]
	Tahun	2023
	Penulis	Maulana Muhammad Jogo Samodro, Muhammad Kunta Biddinika, dan Abdul Fadlil
	Metode	<i>Decision Tree</i> dan <i>Naïve Bayes</i>
	Hasil Penelitian	Penelitian ini membandingkan dua algoritma untuk klasifikasi diabetes, hasil <i>Decision Tree</i> dengan akurasi 96,36%, sedangkan <i>Naïve Bayes</i> mencapai akurasi 90,45%.
9	Jurnal	JATI
	Judul	Evaluation of the Performance of Artificial Neural Network Algorithms with Machine Learning Algorithms in Diabetes Classification [23]
	Tahun	2025
	Penulis	Jojor Putri Pasaribu, Zulfahmi Indra
	Metode	<i>Random Forest</i> , <i>Decision Tree</i> , KNN, ANN, SVM
	Hasil Penelitian	Penelitian ini melakukan komparasi performa ANN dengan beberapa algoritma <i>machine learning</i> . Hasil evaluasi menunjukkan <i>Random Forest</i> 97,58%, <i>Decision Tree</i> 95,16%, KNN 91,13%, ANN 83,87%, dan SVM 82,26%.

Jurnal Terdahulu		
10	Jurnal	Journal of Diabetes Research (Hindawi)
	Judul	Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation [24]
	Tahun	2023
	Penulis	Antonio García Domínguez, Carlos E. Galván Tejada, Rafael Magallanes Quintanar, Hamurabi Gamboa Rosales, Irma González Curiel, Jesús Peralta Romero, dan Miguel Cruz
	Metode	SVM, <i>Random Forest</i> , kNN, <i>Gradient Boosting</i> , <i>Extra Trees</i> , <i>Naïve Bayes</i>
	Hasil Penelitian	Penelitian ini membangun model deteksi diabetes pada pasien Meksiko dengan menggabungkan seleksi fitur <i>Akaike Information Criterion</i> dan <i>Genetic Algorithm</i> , serta membandingkan 6 <i>classifier</i> yaitu SVM, <i>Random Forest</i> , kNN, <i>Gradient Boosting</i> , <i>Extra Trees</i> , <i>Naïve Bayes</i> . Performa yang dilaporkan umumnya 94%, dengan akurasi terbaik mencapai 96% pada kombinasi fitur hasil AIC dengan <i>Random Forest</i> , sementara beberapa model lain berada di kisaran 0,94 sampai 0,95 tergantung skenario subset fitur yang digunakan.

Jurnal pertama digunakan sebagai pijakan utama untuk menegaskan bahwa pemilihan *Artificial Neural Network* (ANN) dalam penelitian ini memiliki dasar akademik yang kuat, karena ANN telah terbukti relevan untuk pola data klinis tabular seperti Pima. Keterkaitannya dengan penelitian ini terletak pada penguatan argumen bahwa model ANN perlu dibangun dengan rancangan pelatihan yang tepat agar tidak mudah mengalami *overfitting*, sehingga penelitian ini menempatkan *pre-processing* dan evaluasi sebagai bagian penting dalam memastikan hasil klasifikasi yang dapat dipertanggungjawabkan.

Jurnal kedua digunakan untuk menempatkan penelitian ini pada konteks klinis yang lebih luas, khususnya ketika literatur membahas klasifikasi tipe diabetes dan kondisi *prediabetes*. Hubungan langsungnya dengan penelitian ini adalah sebagai dasar penegasan ruang lingkup, meskipun literatur dapat mengarah pada klasifikasi multi-kelas, penelitian ini secara metodologis mengikuti struktur label pada dataset yang digunakan, yaitu keluaran biner, sehingga pembahasan tipe diabetes ditempatkan sebagai konteks dan peluang pengembangan, bukan target eksperimen utama.

Jurnal ketiga menjadi penguat bahwa dataset Pima layak diperlakukan sebagai *benchmark* penelitian komparatif, karena literatur juga menguji beberapa algoritma pada dataset yang sama. Keterkaitannya dengan penelitian ini tampak pada pendekatan komparatif yang sejalan, yakni menilai performa beberapa metode pada data yang identik, sehingga pemilihan ANN dan *Decision Tree* pada penelitian ini dapat diposisikan sebagai fokus perbandingan dua pendekatan yang mewakili keseimbangan antara kemampuan prediksi dan keterjelasan keputusan.

Jurnal keempat menjadi rujukan penting untuk menguatkan alasan pemilihan *Decision Tree* dalam penelitian ini, terutama dari sisi interpretabilitas. Hubungannya dengan penelitian ini adalah penegasan bahwa *Decision Tree* relevan ketika penelitian tidak hanya mengejar performa, tetapi juga membutuhkan model yang lebih mudah ditelusuri dan dijelaskan sebagai aturan keputusan, sehingga hasil klasifikasi dapat dikomunikasikan dengan lebih jelas pada pembaca non-teknis.

Jurnal kelima digunakan untuk memperkuat argumen bahwa ANN tidak hanya dipilih sebagai perbandingan, melainkan sebagai model yang memang sering diarahkan untuk kebutuhan klasifikasi medis yang menuntut kecepatan dan ketepatan. Keterkaitannya dengan penelitian ini terletak pada penekanan bahwa performa ANN sangat dipengaruhi oleh strategi pelatihan, sehingga penelitian ini menegaskan pentingnya *pipeline* yang sistematis sejak *pre-processing*, pembagian data, hingga evaluasi agar ANN dapat bekerja stabil pada data klinis tabular.

Jurnal keenam dipakai untuk memposisikan penelitian ini terhadap perkembangan metode yang lebih mutakhir seperti *deep learning*, tanpa harus

menjadikan kompleksitas tersebut sebagai fokus utama. Hubungannya dengan penelitian ini adalah sebagai pembanding konseptual, ketika literatur menunjukkan ruang pengembangan model yang lebih kompleks, penelitian ini memilih fokus yang lebih terukur pada ANN dan *Decision Tree* agar proses analisis tetap transparan, mudah direplikasi, dan sesuai dengan karakter data tabular yang digunakan.

Jurnal ketujuh berfungsi sebagai acuan bahwa peningkatan performa sering dicapai melalui pendekatan *hybrid* atau *ensemble* yang memadukan beberapa komponen model. Keterkaitannya dengan penelitian ini terletak pada penegasan posisi kontribusi, penelitian ini tidak diarahkan untuk membangun model paling kompleks, melainkan menyajikan perbandingan yang jelas antara ANN dan *Decision Tree* sebagai dua pendekatan representatif, serta membuka ruang saran bahwa pengembangan berikutnya dapat mempertimbangkan model hibrida apabila tersedia data yang lebih besar dan kebutuhan implementasi yang lebih luas.

Jurnal kedelapan digunakan untuk memperkuat landasan teoritis bahwa data klinis sering memuat hubungan yang tidak linear dan saling berkaitan antarfaktor, sehingga model berbasis jaringan seperti ANN relevan untuk menangkap pola tersebut. Hubungannya dengan penelitian ini adalah memberi justifikasi bahwa pemilihan ANN memiliki dasar konseptual yang kuat, sementara *Decision Tree* tetap dipertahankan sebagai pembanding yang lebih mudah dijelaskan, sehingga diskusi penelitian ini dapat menyoroti *trade-off* antara kekuatan pemodelan dan keterbacaan keputusan.

Jurnal kesembilan dijadikan rujukan metodologis untuk menekankan bahwa penilaian performa model perlu dilakukan dengan rancangan evaluasi yang kuat, misalnya melalui *hyperparameter tuning* dan *cross-validation*, agar hasil lebih stabil terhadap variasi data. Keterkaitannya dengan penelitian ini adalah sebagai penguat bahwa eksperimen komparatif harus dijalankan pada prosedur evaluasi yang konsisten, serta sebagai landasan untuk menyatakan bahwa peningkatan reliabilitas hasil dapat menjadi arah pengembangan lanjutan ketika penelitian diperluas pada skenario data yang lebih beragam.

Jurnal kesepuluh berperan untuk menghubungkan penelitian ini dengan kebutuhan validasi pada konteks lokal, karena penelitian tersebut menggunakan data klinis dari Indonesia. Keterkaitannya dengan penelitian ini terletak pada penegasan keterbatasan generalisasi dataset publik dan pentingnya pengujian ulang pada data yang merepresentasikan populasi setempat apabila model akan diarahkan ke penerapan nyata, sehingga penelitian ini dapat diposisikan sebagai studi komparatif berbasis data sekunder yang kuat secara metodologis sekaligus menyiapkan jalur uji lanjut pada data klinis Indonesia.

Secara keseluruhan, tinjauan dari sepuluh jurnal yang dianalisis menunjukkan adanya kemajuan yang sangat signifikan dalam pemanfaatan *Machine Learning (ML)* dan *Artificial Intelligence (AI)*, khususnya *Artificial Neural Network (ANN)* dan *Deep Learning (DL)*, dalam bidang diagnosis dan prediksi penyakit diabetes. Seluruh penelitian berangkat dari permasalahan mendasar yang sama, yakni tingginya prevalensi diabetes di dunia dan keterbatasan metode diagnosis konvensional yang masih membutuhkan waktu lama, biaya tinggi, serta ketergantungan besar pada tenaga medis. Melalui pendekatan komputasional, para peneliti berupaya menciptakan sistem diagnosis otomatis yang mampu memproses data medis dalam jumlah besar secara efisien dan menghasilkan prediksi yang akurat. Dari hasil studi, terlihat bahwa penggunaan *Pima Indian Diabetes Dataset (PIDD)* dan data klinis lokal menjadi standar umum karena memiliki parameter fisiologis yang penting seperti kadar glukosa, tekanan darah, BMI, usia, serta riwayat keluarga.

Dari sisi metodologis, berbagai pendekatan *Machine Learning* telah diterapkan, mulai dari model klasik seperti *Decision Tree (C4.5)*, *Random Forest (RF)*, dan *Support Vector Machine (SVM)*, hingga model canggih berbasis *Deep Neural Network (DNN)* dan *Convolutional Neural Network (CNN)*. Model *ensemble* seperti *Random Forest*, *XGBoost*, dan *LightGBM* terbukti memberikan performa yang lebih stabil karena mampu menggabungkan hasil prediksi dari beberapa model dasar untuk meningkatkan akurasi dan mengurangi variansi data. Sementara itu, pendekatan *Deep Learning* menunjukkan keunggulan dalam menangani data medis yang kompleks, baik berupa data numerik maupun citra medis seperti retina atau

hasil pemeriksaan laboratorium. Beberapa penelitian, seperti yang dilakukan oleh jurnal terdahulu, bahkan mengintegrasikan CNN dan XGBoost untuk membentuk model hibrida dengan akurasi mencapai lebih dari 96%, menandakan adanya pergeseran paradigma dari model prediksi sederhana menuju sistem diagnosis cerdas berbasis *hybrid intelligence*.

Selain itu, hasil komparatif antar penelitian juga menunjukkan bahwa peningkatan performa tidak hanya dipengaruhi oleh algoritma, tetapi juga oleh kualitas data, teknik prapemrosesan, serta parameter optimasi. Beberapa studi menyoroti pentingnya *learning rate*, *regularization*, dan *optimization algorithm* seperti *Levenberg–Marquardt* atau *Scaled Conjugate Gradient* untuk meningkatkan stabilitas pelatihan ANN. Penelitian lainnya, menunjukkan pentingnya analisis faktor dominan. Misalnya *polyuria* dan *polydipsia* yang secara langsung memengaruhi risiko diabetes. Dengan demikian, kombinasi antara pemilihan fitur yang tepat (*feature selection*) dan model pembelajaran yang adaptif menjadi kunci keberhasilan sistem prediksi berbasis AI.

Secara empiris, seluruh penelitian menunjukkan hasil akurasi yang tinggi, umumnya berada pada rentang 90%–96%, dengan tingkat *precision*, *recall*, dan *F1-score* yang juga cukup stabil. Hal ini menegaskan bahwa pendekatan berbasis *Machine Learning* tidak hanya layak digunakan untuk penelitian, tetapi juga dapat diimplementasikan pada sistem pendukung keputusan medis (*Decision Support System*) di fasilitas kesehatan. Beberapa studi bahkan telah mengusulkan aplikasi berbasis web atau perangkat lunak cerdas yang mampu memberikan hasil diagnosis otomatis hanya dengan memasukkan data pasien. Penerapan sistem ini sangat relevan bagi negara berkembang seperti Indonesia, di mana keterbatasan tenaga medis menjadi tantangan utama dalam upaya deteksi dini penyakit kronis.

Dari perspektif ilmiah, tinjauan ini memperlihatkan adanya pergeseran signifikan dalam penelitian medis menuju era *data-driven healthcare*, di mana pengambilan keputusan klinis tidak lagi semata bergantung pada intuisi dokter, tetapi juga pada hasil analisis cerdas berbasis data. Model *hybrid AI*, *ensemble learning*, serta *deep learning architecture* menjadi tren utama yang terus

berkembang karena mampu menangani data besar, heterogen, dan kompleks dengan efisiensi tinggi. Di sisi lain, tantangan utama ke depan adalah meningkatkan *interpretabilitas* model AI agar hasil prediksi dapat lebih mudah dipahami oleh praktisi medis, serta memastikan keamanan dan privasi data pasien.

Dengan demikian, dapat disimpulkan bahwa sepuluh penelitian ini secara kolektif memberikan kontribusi besar terhadap pengembangan sistem diagnosis diabetes yang lebih akurat, cepat, dan efisien. Integrasi antara algoritma cerdas, data klinis yang berkualitas, serta optimasi komputasional membuka peluang besar bagi penerapan sistem *Artificial Intelligence* dalam dunia medis modern. Tren ini menunjukkan bahwa masa depan deteksi dini dan pencegahan diabetes akan semakin bergantung pada teknologi cerdas yang mampu belajar dan beradaptasi dari data pasien secara berkelanjutan.

2.2 Teori Penelitian

2.2.1 Diabetes Mellitus

Diabetes Mellitus merupakan salah satu penyakit tidak menular yang paling banyak diderita di seluruh dunia dan menjadi masalah kesehatan masyarakat global. Penyakit ini ditandai dengan meningkatnya kadar glukosa dalam darah atau *hyperglycemia* akibat gangguan pada sekresi maupun fungsi hormon insulin. Diabetes menempati peringkat keempat penyebab kematian di dunia dengan jumlah penderita yang terus meningkat setiap tahunnya. Diabetes secara umum dikategorikan menjadi tiga tipe utama, yaitu tipe 1, tipe 2, dan *gestational diabetes mellitus* (GDM). Tipe 1 disebabkan oleh kerusakan sel β pankreas yang menyebabkan tubuh tidak mampu memproduksi insulin. Tipe 2 merupakan jenis yang paling banyak ditemukan, yaitu ketika tubuh mengalami resistensi terhadap insulin sehingga tidak dapat memanfaatkannya secara efektif. Sedangkan *gestational diabetes* terjadi pada masa kehamilan dan umumnya bersifat sementara, namun dapat meningkatkan risiko terjadinya diabetes tipe 2 di kemudian hari [25].

Faktor risiko yang berperan dalam perkembangan penyakit diabetes melitus antara lain usia, riwayat keluarga, obesitas, tekanan darah tinggi, serta gaya

hidup tidak sehat seperti kurang aktivitas fisik dan pola makan tinggi gula. Dampak jangka panjang dari penyakit ini sangat kompleks, di antaranya dapat menyebabkan komplikasi berat seperti penyakit jantung koroner, kerusakan ginjal (*nephropathy*), kebutaan (*retinopathy*), dan gangguan saraf (*neuropathy*). Oleh karena itu, deteksi dini dan klasifikasi pasien berisiko diabetes menjadi langkah penting untuk mengurangi angka morbiditas dan mortalitas akibat penyakit ini.

Dalam konteks teknologi informasi, khususnya *machine learning*, data medis pasien diabetes sering digunakan sebagai objek penelitian karena memiliki pola data yang kompleks namun terukur. Dataset seperti *Pima Indians Diabetes Dataset* menjadi populer karena mengandung berbagai parameter medis seperti kadar glukosa, tekanan darah, jumlah kehamilan, indeks massa tubuh (BMI), kadar insulin, serta usia pasien. Atribut-atribut tersebut dapat digunakan untuk membangun model klasifikasi guna mengelompokkan pasien menjadi dua kategori, yaitu positif atau negatif diabetes, berdasarkan hubungan antarvariabel medis yang terdapat pada data tersebut.

2.2.2 Konsep Data Mining dan Machine Learning

Data mining merupakan proses penemuan pola atau pengetahuan baru dari kumpulan data besar dengan menggunakan teknik statistik, matematika, serta kecerdasan buatan. Dalam dunia akademik dan industri, *data mining* menjadi bagian penting dari bidang yang lebih luas yaitu *Knowledge Discovery in Databases (KDD)*. Proses ini mencakup tahapan seperti pembersihan data, integrasi data, seleksi, transformasi, penggalian pola, evaluasi pengetahuan, dan penyajian hasil. Salah satu teknik paling populer dalam *data mining* adalah *classification*, yaitu proses pengelompokan data ke dalam kategori tertentu berdasarkan fitur atau karakteristik yang telah diketahui sebelumnya.

Sementara itu, *machine learning* adalah cabang dari *artificial intelligence* yang berfokus pada pengembangan algoritma yang memungkinkan komputer untuk belajar dari data tanpa pemrograman eksplisit. Dalam konteks penelitian ini, *machine learning* digunakan untuk mengembangkan model klasifikasi

penyakit diabetes yang mampu mengenali pola dari data medis pasien. Model yang dibangun akan belajar dari data latih (*training data*) untuk memahami hubungan antaratribut dan kemudian diuji menggunakan data uji (*testing data*) guna mengukur performa klasifikasinya. Pendekatan ini menjadikan *machine learning* sangat efektif dalam bidang kesehatan, karena mampu mengidentifikasi pola-pola yang tidak terlihat secara langsung oleh tenaga medis.

2.2.3 Konsep Klasifikasi dalam Machine Learning

Klasifikasi merupakan salah satu teknik utama dalam *supervised learning*, yaitu pendekatan pembelajaran mesin di mana setiap data masukan telah diberi label kelas yang benar. Tujuan utama dari proses klasifikasi adalah untuk mempelajari fungsi atau model yang dapat memetakan input ke dalam kelas yang sesuai. Dalam konteks diagnosis penyakit, klasifikasi berfungsi untuk mengelompokkan pasien berdasarkan atribut medis menjadi “positif diabetes” atau “negatif diabetes”.

Algoritma klasifikasi bekerja dengan menganalisis data berlabel untuk membangun fungsi pemisah antara dua atau lebih kelas. Model yang dihasilkan kemudian digunakan untuk memprediksi kelas data baru yang belum pernah dilihat sebelumnya. Beberapa metrik yang umum digunakan untuk mengevaluasi hasil klasifikasi antara lain *accuracy* (tingkat ketepatan prediksi), *precision* (tingkat ketepatan hasil positif), *recall* atau *sensitivity* (kemampuan mendekripsi kasus positif), dan *F1-score* (rata-rata harmonik dari *precision* dan *recall*). Dalam penelitian medis, metrik-metrik ini sangat penting karena ketidakseimbangan antara data positif dan negatif dapat memengaruhi kualitas model secara signifikan.

2.2.4 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) merupakan algoritma pembelajaran mesin yang terinspirasi oleh cara kerja jaringan saraf biologis manusia. ANN terdiri dari unit-unit kecil yang disebut neuron, yang terhubung satu sama lain melalui bobot (*weights*). Setiap neuron menerima input, mengalikannya dengan bobot

tertentu, menambahkan bias, dan kemudian menerapkan fungsi aktivasi untuk menghasilkan output. Secara umum, struktur ANN terdiri dari tiga lapisan utama yaitu *input layer*, *hidden layer*, dan *output layer*.

Kelebihan utama dari ANN adalah kemampuannya untuk mengenali hubungan non-linear antara variabel input dan output. Hal ini sangat penting dalam analisis data medis yang kompleks seperti data diabetes, di mana hubungan antarvariabel tidak selalu linier atau mudah dijelaskan dengan metode statistik konvensional [26]. Proses pelatihan ANN dilakukan dengan algoritma *backpropagation*, di mana kesalahan hasil prediksi dikembalikan ke jaringan untuk memperbarui bobot hingga diperoleh kesalahan minimum. Fungsi aktivasi yang sering digunakan dalam penelitian klasifikasi antara lain *sigmoid*, *ReLU (Rectified Linear Unit)*, dan *tanh*.

Namun demikian, ANN juga memiliki kelemahan, seperti kebutuhan komputasi yang tinggi dan interpretasi hasil yang sulit. Dalam konteks medis, meskipun ANN menghasilkan akurasi tinggi, tenaga kesehatan sering kali kesulitan memahami bagaimana model mencapai keputusan tertentu karena sifatnya yang menyerupai *black box*.

2.2.5 Decision Tree

Decision Tree merupakan algoritma klasifikasi berbasis struktur pohon yang bekerja dengan memecah dataset menjadi subset yang lebih kecil berdasarkan nilai atribut tertentu. Setiap percabangan dalam pohon keputusan mewakili kondisi logis pada atribut, sedangkan daun (*leaf node*) menunjukkan hasil klasifikasi. Algoritma populer seperti *ID3*, *C4.5*, dan *CART* (Classification and Regression Tree) sering digunakan untuk membangun *Decision Tree*.

Kelebihan *Decision Tree* adalah kemampuannya untuk menghasilkan model yang mudah dipahami dan divisualisasikan, sehingga cocok digunakan dalam bidang medis yang membutuhkan transparansi hasil. Namun, DT cenderung rentan terhadap *overfitting*, terutama ketika dataset mengandung banyak variabel atau noise. Oleh karena itu, proses *pruning* atau pemangkasan pohon sering dilakukan untuk meningkatkan generalisasi model [27].

2.2.6 Metodologi CRISP-DM

Metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan kerangka kerja standar internasional yang digunakan untuk mengembangkan proyek *data mining* secara sistematis. CRISP-DM terdiri dari enam tahap utama, yaitu *Problem Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*.

Pada tahap *Problem Understanding*, penulis mengidentifikasi tujuan penelitian dan kebutuhan analisis yang akan dilakukan. Tahap *Data Understanding* melibatkan eksplorasi awal dataset untuk mengenali struktur, ukuran, dan karakteristik data. *Data Preparation* mencakup proses pembersihan, transformasi, dan pembagian dataset menjadi data latih dan data uji. *Modeling* adalah tahap pembangunan model menggunakan algoritma klasifikasi, dalam hal ini *Artificial Neural Network* dan *Decision Tree*. Setelah model terbentuk, tahap *Evaluation* dilakukan untuk menilai performa model berdasarkan metrik evaluasi. Terakhir, tahap *Deployment* digunakan untuk menerapkan hasil model ke dalam sistem atau menghasilkan laporan yang dapat dimanfaatkan dalam pengambilan keputusan [28].

CRISP-DM dianggap sangat efektif karena bersifat fleksibel, independen dari perangkat lunak tertentu, dan dapat digunakan untuk berbagai jenis data, termasuk data medis. Dalam penelitian ini, metode CRISP-DM digunakan untuk memastikan bahwa proses klasifikasi data diabetes berjalan terstruktur, terukur.

2.3 Framework/Algoritma yang digunakan

Penelitian ini menggunakan pendekatan *machine learning* dengan menerapkan dua algoritma klasifikasi utama, yaitu *Artificial Neural Network* (ANN) dan *Decision Tree*. Kedua algoritma ini dipilih karena mewakili dua pendekatan berbeda dalam proses klasifikasi data yaitu ANN merupakan metode berbasis jaringan saraf tiruan yang meniru cara kerja sistem biologis otak manusia, sedangkan *Decision Tree* merupakan metode berbasis aturan (*rule-based*) yang membangun model keputusan dalam bentuk struktur pohon.

Pemilihan kedua algoritma ini bertujuan untuk melakukan perbandingan kinerja dari dua paradigma pembelajaran yang berbeda seperti pendekatan *connectionist* (ANN) dan pendekatan *symbolic learning (Decision Tree)*. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan gambaran empiris mengenai kelebihan dan keterbatasan masing-masing algoritma dalam mengklasifikasikan data medis pasien diabetes.

2.3.1 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) merupakan salah satu algoritma *supervised learning* yang terinspirasi dari sistem saraf manusia, di mana setiap neuron berfungsi sebagai unit pemrosesan informasi. ANN dibangun dari sejumlah lapisan neuron yang saling terhubung melalui bobot (*weights*) dan bias. Setiap neuron menerima input, melakukan perhitungan linier dengan bobot tertentu, kemudian menerapkan fungsi aktivasi (*activation function*) untuk menghasilkan output non-linear.

Proses pembelajaran ANN dilakukan menggunakan algoritma *backpropagation*, yaitu algoritma yang mengoptimalkan bobot jaringan dengan meminimalkan kesalahan prediksi (*loss function*). Kesalahan yang dihasilkan dari selisih antara output aktual dan output yang diprediksi dihitung menggunakan fungsi error seperti *Mean Squared Error (MSE)* atau *Binary Cross-Entropy*, tergantung pada jenis klasifikasinya. Nilai error tersebut kemudian digunakan untuk memperbarui bobot dengan metode *gradient descent* atau turunannya seperti *Adam Optimizer*.

Kelebihan utama ANN adalah kemampuannya mempelajari hubungan non-linear antarvariabel serta keakuratannya yang tinggi untuk dataset kompleks. Namun, ANN juga memiliki kekurangan, yaitu proses pelatihan yang relatif lama, ketergantungan tinggi pada ukuran data, serta sulitnya interpretasi hasil karena sifatnya yang menyerupai *black box*. Dalam konteks penelitian medis, interpretasi ini menjadi penting karena hasil klasifikasi harus dapat dijelaskan kepada tenaga profesional kesehatan.

Dalam penelitian ini, ANN diimplementasikan menggunakan pustaka *TensorFlow* dan *Keras* dalam bahasa pemrograman *Python*. Parameter seperti *epoch*, *batch size*, dan *learning rate* disesuaikan agar diperoleh hasil klasifikasi yang optimal tanpa menyebabkan *overfitting*. Model kemudian dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

2.3.2 Decision Tree

Decision Tree adalah algoritma klasifikasi yang bekerja dengan membangun struktur berbentuk pohon keputusan dari dataset. Setiap simpul (*node*) pada pohon merepresentasikan atribut yang diuji, setiap cabang menunjukkan hasil dari pengujian, dan setiap daun (*leaf node*) menunjukkan kelas hasil klasifikasi. Tujuan utama dari *Decision Tree* adalah mempartisi data ke dalam subset yang homogen berdasarkan nilai atribut yang paling berpengaruh terhadap target klasifikasi.

Algoritma Decision Tree membangun pohon secara rekursif menggunakan prinsip *divide and conquer*. Pada setiap tahap pemisahan (*splitting*), algoritma akan memilih atribut terbaik menggunakan ukuran seperti *Information Gain*, *Entropy*, atau *Gini Index*.

Kelebihan utama *Decision Tree* adalah kemudahan interpretasi hasil karena struktur pohon yang dihasilkan dapat divisualisasikan dan dipahami oleh pengguna non-teknis, termasuk dokter atau analis medis. Selain itu, DT juga memiliki waktu komputasi yang relatif cepat dan tidak memerlukan normalisasi data yang kompleks. Namun, DT cenderung mengalami masalah *overfitting*, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data latih sehingga kurang mampu menggeneralisasi data baru. Untuk mengatasinya, digunakan teknik *pruning* atau pemangkasan cabang pohon yang tidak signifikan.

Dalam penelitian ini, algoritma *Decision Tree* diimplementasikan menggunakan pustaka *Scikit-Learn* dalam bahasa pemrograman *Python*, dengan variasi model *CART* (*Classification and Regression Tree*). Model

dievaluasi menggunakan metrik yang sama seperti pada ANN agar perbandingan performa kedua algoritma dapat dilakukan secara objektif.

2.3.3 Kerangka CRISP-DM dalam Penerapan Klasifikasi

Metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) digunakan sebagai kerangka kerja utama dalam penelitian ini. CRISP-DM dipilih karena bersifat sistematis, terstruktur, dan dapat diterapkan di berbagai domain termasuk bidang kesehatan. Proses penelitian berdasarkan CRISP-DM terdiri dari enam tahap utama, yaitu:

1. Problem Understanding

Tahap ini berfokus pada pemahaman masalah yang ingin diselesaikan, yaitu bagaimana mengklasifikasikan data pasien diabetes secara akurat menggunakan dua algoritma berbeda. Penulis menetapkan tujuan penelitian berupa analisis komparatif antara ANN dan *Decision Tree* untuk menentukan model klasifikasi terbaik.

2. Data Understanding

Pada tahap ini dilakukan eksplorasi terhadap Pima Indians Diabetes Dataset guna memahami struktur, jumlah atribut, distribusi kelas, serta karakteristik data. Analisis statistik deskriptif dan visualisasi data dilakukan untuk mendeteksi *missing values, outliers, dan class imbalance*.

3. Data Preparation

Tahap ini meliputi pembersihan data (*data cleaning*), penanganan nilai kosong, normalisasi data, dan pembagian dataset menjadi training set dan testing set dengan rasio tertentu (misalnya 80:20). Proses *feature scaling* dilakukan agar algoritma ANN dapat bekerja optimal, sedangkan *Decision Tree* tidak terlalu bergantung pada skala data.

4. Modeling

Pada tahap ini dilakukan pembangunan dua model klasifikasi berbeda, yaitu model ANN dan *Decision Tree*. Parameter seperti

jumlah *neuron*, fungsi aktivasi, dan kedalaman pohon (*max depth*) diuji untuk menemukan kombinasi parameter terbaik.

5. Evaluation

Tahap ini bertujuan mengevaluasi performa model dengan metrik kuantitatif seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Hasil evaluasi kedua model dibandingkan untuk menentukan algoritma mana yang paling efektif dalam klasifikasi diabetes.

6. Deployment (opsional)

Tahap akhir ini berfokus pada dokumentasi hasil, pembuatan visualisasi performa model, serta interpretasi hasil klasifikasi. Hasil akhir dapat dijadikan acuan bagi penelitian lanjutan atau pengembangan sistem diagnosis berbasis machine learning di bidang kesehatan. Namun hal ini opsional dikarenakan penelitian ini menggunakan data sekunder anonim *Pima Indians Diabetes*, yang tidak berasal dari lingkungan operasional nyata seperti rumah sakit atau klinik.

Dengan menerapkan kerangka CRISP-DM, seluruh proses penelitian menjadi lebih transparan, sistematis, dan dapat direplikasi oleh penulis lain. Framework ini juga memastikan bahwa hasil klasifikasi tidak hanya akurat secara matematis, tetapi juga relevan dengan konteks medis yang menjadi fokus penelitian.

2.4 Tools/software yang digunakan

Dalam penelitian ini, pemilihan *tools* dan *software* memiliki peranan yang sangat penting dalam mendukung seluruh tahapan proses penelitian, mulai dari eksplorasi data hingga evaluasi model klasifikasi. Agar proses penelitian berjalan secara sistematis dan efisien, digunakan kombinasi antara bahasa pemrograman *Python* dan lingkungan kerja interaktif *Jupyter Notebook*, yang keduanya saling melengkapi dalam penerapan metode *CRISP-DM (Cross-Industry Standard Process for Data Mining)*.

Bahasa pemrograman *Python* digunakan karena merupakan bahasa yang paling umum digunakan dalam bidang *data science* dan *machine learning*. *Python* dikenal sebagai bahasa pemrograman tingkat tinggi yang bersifat *open-source*, fleksibel, mudah dipahami, serta memiliki komunitas pengguna yang luas. Keunggulan utama *Python* terletak pada ekosistem pustaka (*library*) yang sangat kaya untuk analisis data, seperti *Pandas*, *NumPy*, *Scikit-Learn*, *TensorFlow*, dan *Matplotlib*. Dengan pustaka-pustaka tersebut, *Python* memungkinkan penulis untuk melakukan seluruh tahapan *CRISP-DM* secara terintegrasi, mulai dari *data understanding*, *data preparation*, *modeling*, hingga *evaluation*.

Salah satu lingkungan kerja yang paling umum digunakan dalam *Python* adalah *Jupyter Notebook*. *Jupyter Notebook* merupakan *open-source interactive environment* yang memungkinkan pengguna untuk menulis dan menjalankan kode secara interaktif dalam format sel (*cell-based environment*). Setiap sel dapat berisi kode, teks deskriptif, visualisasi, ataupun hasil keluaran (*output*). Kelebihan utama dari *Jupyter Notebook* adalah kemampuannya menggabungkan kode pemrograman, dokumentasi teks, dan visualisasi data dalam satu berkas kerja yang terstruktur dan mudah dipahami. Hal ini sangat membantu dalam konteks penelitian akademik, karena seluruh proses analisis dapat terdokumentasi dengan baik dan dapat direplikasi oleh penulis lain.

Selain itu, *Jupyter Notebook* juga mendukung integrasi langsung dengan berbagai pustaka visualisasi data seperti *Matplotlib* dan *Seaborn*, yang digunakan dalam penelitian ini untuk menampilkan grafik distribusi data, hubungan antaratribut, serta hasil evaluasi model dalam bentuk *confusion matrix* dan *ROC curve*. Visualisasi semacam ini sangat penting untuk memahami performa algoritma klasifikasi dan pola distribusi data pasien diabetes pada *Pima Indians Diabetes Dataset*.

Pustaka *Pandas* berfungsi sebagai alat utama untuk manipulasi data tabular. Dengan *Pandas*, penulis dapat melakukan pembersihan data (*data cleaning*), penghapusan nilai kosong (*missing values*), transformasi atribut, serta analisis statistik deskriptif dengan efisien. Pustaka *NumPy* digunakan sebagai fondasi

numerik dalam komputasi ilmiah, terutama untuk pengolahan array dan operasi matematis tingkat tinggi seperti perhitungan matriks dan vektor. Dalam penelitian ini, *NumPy* digunakan untuk mendukung perhitungan matematis pada algoritma *Artificial Neural Network* dalam proses propagasi maju (*forward propagation*) dan pembaruan bobot (*backpropagation*).

Pustaka *Scikit-Learn* digunakan untuk membangun model *Decision Tree* serta melakukan pembagian data menjadi data latih (*training set*) dan data uji (*testing set*). Pustaka ini juga menyediakan fungsi untuk normalisasi data, pembentukan model klasifikasi, serta perhitungan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Dengan pustaka ini, penulis dapat melakukan pengujian model dengan cepat, akurat, dan mudah direplikasi.

Untuk implementasi *Artificial Neural Network*, digunakan pustaka *TensorFlow* dan *Keras*. *TensorFlow* merupakan platform *machine learning* yang dikembangkan oleh Google dan banyak digunakan dalam penelitian akademik maupun industri. *Keras*, sebagai antarmuka tingkat tinggi dari *TensorFlow*, digunakan untuk mendefinisikan struktur jaringan saraf tiruan secara lebih intuitif. Dalam penelitian ini, *Keras* digunakan untuk menentukan jumlah neuron, jumlah *hidden layer*, fungsi aktivasi seperti *sigmoid* dan *ReLU (Rectified Linear Unit)*, serta pengaturan parameter pelatihan seperti *learning rate* dan *epoch*.

Jupyter Notebook juga memungkinkan dokumentasi dan reproduksibilitas hasil penelitian. Setiap langkah analisis dapat dijalankan ulang kapan saja tanpa kehilangan konteks atau urutan penggerjaan. Ini menjadikan *Jupyter Notebook* sebagai alat yang ideal untuk implementasi penelitian berbasis *machine learning* seperti klasifikasi data medis. Selain itu, lingkungan kerja ini sangat mendukung proses kolaborasi karena hasil penelitian dapat diekspor ke berbagai format seperti *HTML*, *PDF*, atau *Markdown*, yang memudahkan proses pelaporan hasil ke dosen pembimbing maupun publikasi ilmiah.