# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Perkembangan teknologi informasi dan *artificial intelligence* (AI) pada dekade terakhir telah menghadirkan kemajuan signifikan dalam dunia kesehatan. Salah satu cabang AI yang berkembang pesat adalah *machine learning*, yaitu pendekatan komputasional yang memungkinkan sistem komputer mempelajari pola dari data dan melakukan analisis atau pengambilan keputusan tanpa pemrograman eksplisit. Dalam konteks medis, *machine learning* memiliki potensi besar dalam membantu diagnosis dini, klasifikasi penyakit, dan analisis risiko kesehatan pasien berdasarkan data klinis [1].

Salah satu penyakit yang menjadi perhatian global adalah diabetes melitus, yaitu gangguan metabolik kronis yang ditandai oleh peningkatan kadar glukosa darah akibat gangguan sekresi atau kerja hormon insulin. Menurut laporan World Health Organization, diabetes telah menjadi penyebab utama morbiditas dan mortalitas di seluruh dunia. Diketahui bahwa sekitar 537 juta orang dewasa di dunia hidup dengan diabetes, dan angka ini diperkirakan akan terus meningkat menjadi 643 juta pada tahun 2030. Di sisi lain, diabetes tidak hanya menimbulkan beban klinis, tetapi juga beban ekonomi. Pengobatan jangka panjang, kontrol rutin, pemeriksaan laboratorium, serta risiko rawat inap saat komplikasi muncul sering membuat biaya perawatan menjadi tinggi, baik bagi individu maupun sistem kesehatan. Karena itu, upaya pencegahan dan deteksi dini menjadi penting. Semakin cepat risiko dapat diidentifikasi, semakin besar peluang dilakukan intervensi gaya hidup dan penanganan awal agar komplikasi dapat ditekan [2].

Dalam konteks gangguan metabolik, dikenal dengan prediabetes yaitu merupakan kondisi kadar gula darah sudah berada di atas normal tetapi belum memenuhi kriteria diabetes. Selain itu, secara klinis diabetes terbagi menjadi beberapa tipe, terutama diabetes tipe 1 dan tipe 2, yang memiliki mekanisme berbeda. Namun, penelitian berbasis dataset sekunder perlu mengikuti struktur label

yang tersedia agar kesimpulan tidak melampaui data. Oleh sebab itu, penelitian ini tidak berfokuskan untuk mengklasifikasi prediabetes maupun tipe diabetes, melainkan fokus pada klasifikasi biner terindikasi diabetes dan tidak diabetes sesuai label pada dataset [3].

Dalam bidang *data mining*, proses klasifikasi merupakan salah satu teknik utama yang digunakan untuk mengelompokkan data ke dalam kelas-kelas tertentu berdasarkan karakteristik atau fitur yang dimilikinya. Dalam konteks penelitian ini, proses klasifikasi bertujuan untuk menentukan apakah seorang pasien termasuk ke dalam kategori "menderita diabetes" atau "tidak menderita diabetes" berdasarkan data medis seperti kadar glukosa, tekanan darah, BMI, kadar insulin, dan usia [4].

Pada sisi lain, ketersediaan dataset medis yang terstruktur membuka peluang penerapan *machine learning* sebagai alat bantu untuk mengklasifikasikan pasien berdasarkan tingkat risiko. Penelitian ini menggunakan data klinis tabular yang merujuk pada dataset Pima Indians Diabetes, yang banyak digunakan dalam studi prediksi diabetes sebagai dataset pembanding [5]. Data yang digunakan mempunyai 1.200 observasi, 8 variabel prediktor, dan 1 variabel target biner (*Outcome*). Prediktor yang digunakan meliputi *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction,* dan *Age*. Variabel Pregnancies di dalam dataset dipahami sebagai riwayat jumlah kehamilan, sehingga penelitian ini bukan penelitian khusus diabetes ibu hamil atau diabetes gestasional, melainkan studi klasifikasi status diabetes pada populasi perempuan dewasa sesuai karakter dataset.

Berbagai penelitian terdahulu telah mengaplikasikan beragam algoritma klasifikasi pada Pima Indians Diabetes Dataset, mulai dari *k-Nearest Neighbor*, *Support Vector Machine*, *Naive Bayes*, *Logistic Regression*, hingga *Artificial Neural Network* dan *Decision Tree*. Hasil yang dilaporkan menunjukkan variasi akurasi dan metrik evaluasi lainnya, bergantung pada pemilihan algoritma, teknik *preprocessing*, dan parameter yang digunakan [6]. Namun demikian, banyak penelitian hanya berfokus pada satu algoritma tertentu atau membandingkan beberapa algoritma dalam kondisi eksperimen yang tidak sepenuhnya seragam,

sehingga sulit ditarik kesimpulan yang konsisten mengenai algoritma mana yang lebih sesuai untuk data medis tabular seperti *Pima Indians Diabetes Dataset*.

Di antara berbagai algoritma yang sering digunakan, *Artificial Neural Network* (ANN) dan *Decision Tree* termasuk dua pendekatan yang menonjol dan memiliki karakteristik yang berbeda. ANN dikenal sebagai model yang mampu mempelajari hubungan *non-linear* yang kompleks antara fitur dan kelas. Dengan struktur jaringan berlapis dan bobot yang dapat dioptimasi, ANN berpotensi memberikan akurasi yang tinggi, tetapi sering dipandang sebagai model *black box* karena sulit dijelaskan secara intuitif kepada pengguna non-teknis. Sebaliknya, *Decision Tree* membangun model dalam bentuk struktur pohon keputusan yang terdiri dari simpul-simpul pemisah dan daun yang merepresentasikan kelas. Struktur ini mudah diterjemahkan menjadi aturan *IF–THEN*, sehingga hasil klasifikasinya relatif mudah dipahami dan dijelaskan [7].

Pemilihan kedua algoritma ini dalam satu penelitian bukan semata-mata karena keduanya populer, tetapi karena ANN dan Decision Tree mewakili dua paradigma yang berbeda dalam *machine learning*. ANN menekankan fleksibilitas dan kemampuan menangkap pola non-linear, sedangkan Decision Tree menekankan interpretabilitas dan transparansi keputusan [8]. Masalah utama penelitian ini terletak pada dua hal. Pertama, dataset klinis tabular sering mengandung nilai yang secara medis kurang masuk akal yaitu nilai 0 pada variabel fisiologis tertentu, sehingga diperlukan *preprocessing* yang tepat agar model tidak belajar dari pola yang keliru. Pada dataset yang digunakan, nilai 0 masih ditemukan pada beberapa fitur seperti *Insulin* dan *SkinThickness*, sehingga perlu ditangani secara sistematis sebelum pemodelan. Kedua, banyak penelitian hanya menyebut membandingkan algoritma tanpa menegaskan alasan pemilihan model dan tanpa menempatkan perbandingan itu pada tujuan praktis yang jelas. Padahal, dalam konteks keputusan berbasis data, kinerja prediksi perlu ditimbang bersama interpretabilitas model agar hasil analisis mendapatkan maksimal.

Selain pemilihan algoritma, tahapan *preprocessing* data juga memegang peran penting dalam menentukan performa model. Pima Indians Diabetes Dataset

memiliki karakteristik khusus, antara lain adanya nilai nol pada atribut yang secara klinis tidak mungkin bernilai nol, sehingga lebih tepat diperlakukan sebagai *missing value*. Jika aspek ini diabaikan, model yang dihasilkan berisiko mengambil kesimpulan yang keliru. Beberapa penelitian terdahulu tidak selalu menjelaskan secara rinci bagaimana penanganan nilai nol, pembagian data latih dan uji, serta proses normalisasi dilakukan. Padahal, keputusan teknis pada tahap *data preparation* sangat mempengaruhi kinerja akhir algoritma [9].

Untuk mengatasi hal tersebut, penelitian ini menerapkan kerangka kerja *CRISP–DM* (*Cross Industry Standard Process for Data Mining*) yang mencakup tahapan *problem understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Penerapan *CRISP–DM* secara eksplisit diharapkan dapat menghasilkan alur kerja yang sistematis, transparan, dan dapat direplikasi oleh peneliti lain [10]. Pada tahap *modeling*, ANN dan *Decision Tree* akan dibangun dan dilatih dengan skema *preprocessing* yang sama, kemudian dievaluasi menggunakan beberapa metrik, tidak hanya akurasi tetapi juga *precision*, *recall*, dan *F1-score*.

Secara teknis, penelitian ini memiliki urgensi karena berupaya mengisi kekosongan dalam penelitian terdahulu yang umumnya menggunakan ANN atau *Decision Tree* secara terpisah, atau membandingkannya dengan algoritma lain dalam kondisi eksperimen yang berbeda. Dengan membandingkan *Artificial Neural Network* dan *Decision Tree* secara langsung pada *Pima Indians Diabetes Dataset* dalam satu kerangka *CRISP–DM* yang terdokumentasi dengan jelas, penelitian ini diharapkan dapat memberikan dasar empiris yang lebih kuat mengenai pemilihan algoritma untuk klasifikasi diabetes pada data medis tabular [11]. Hasilnya tidak hanya menunjukkan model mana yang memiliki kinerja terbaik, tetapi juga memberikan gambaran mengenai keseimbangan antara akurasi, kompleksitas model, dan interpretabilitas, yang penting bagi pengembangan sistem pendukung keputusan diagnosis dini diabetes berbasis *machine learning*.

## 1.2 Rumusan Masalah

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

1. Bagaimana penerapan metode CRISP-DM dalam proses klasifikasi data medis pada *Pima Indians Diabetes Dataset*?
2. Bagaimana kinerja algoritma *Artificial Neural Network* dan *Decision Tree* dalam mengklasifikasikan pasien berdasarkan variabel?
3. Algoritma manakah yang memiliki performa terbaik berdasarkan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*?

## 1.3 Batasan Masalah

Agar penelitian ini terfokus dan dapat diselesaikan dengan baik, maka batasan masalah yang ditetapkan adalah:

1. Data yang digunakan dalam penelitian ini adalah *Pima Indians Diabetes Dataset* dari *National Institute of Diabetes and Digestive and Kidney*.
2. Atribut yang digunakan dalam proses klasifikasi meliputi delapan variable yaitu *Pregnancies*, *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, *BMI*, *Diabetes Pedigree Function*, dan *Age.*
3. Penelitian hanya membandingkan dua algoritma klasifikasi, yaitu *Artificial Neural Network* dan *Decision Tree*.
4. Evaluasi kinerja model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.
5. Penelitian tidak mencakup optimasi parameter lanjutan seperti *grid search* atau *hyperparameter tuning* tingkat lanjut.
6. Implementasi dilakukan menggunakan bahasa pemrograman *Python* dengan pustaka *Scikit-Learn*, *TensorFlow*, *Pandas*, dan *Matplotlib*.

## 1.4 Tujuan dan Manfaat Penelitian

### 1.4.1 Tujuan Penelitian

Adapun tujuan khusus dari penelitian ini adalah sebagai berikut:

1. Menerapkan kerangka kerja *CRISP–DM* pada proses pengolahan dan analisis *Pima Indians Diabetes Dataset* untuk membangun model klasifikasi diabetes.
2. Membangun dan mengevaluasi model klasifikasi diabetes menggunakan algoritma *Artificial Neural Network* dan *Decision Tree* dengan tahapan *preprocessing* dan skema evaluasi yang seragam.
3. Melakukan analisis komparatif terhadap kinerja algoritma Artificial Neural Network dan Decision Tree berdasarkan metrik accuracy, precision, recall, dan F1-score sebagai dasar empiris untuk menjelaskan perbedaan hasil pada penelitian-penelitian sebelumnya yang menggunakan Pima Indians Diabetes Dataset.

### 1.4.2 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan kontribusi nyata, baik dari segi teoritis, praktis, sosial, maupun teknologi. Adapun manfaat yang diharapkan adalah sebagai berikut:

1. Manfaat teoritis: memberikan kontribusi terhadap pengembangan kajian *machine learning* di bidang kesehatan, khususnya terkait perbandingan algoritma *Artificial Neural Network* dan *Decision Tree* pada *Pima Indians Diabetes Dataset*, serta penerapan kerangka kerja *CRISP–DM* dalam pembangunan model klasifikasi medis.
2. Manfaat praktis: menyediakan informasi bagi praktisi kesehatan, analis data, dan pengembang sistem mengenai algoritma yang lebih sesuai untuk digunakan sebagai model klasifikasi risiko diabetes pada data medis tabular, dengan mempertimbangkan aspek akurasi dan interpretabilitas.
3. Manfaat metodologis: menyajikan *pipeline* analisis yang terdokumentasi dengan baik mulai dari *data understanding*, *data preparation*, *modeling*, hingga *evaluation*, yang dapat dijadikan referensi dan contoh kerja bagi penulis atau mahasiswa yang akan melakukan penelitian serupa.

4. Manfaat bagi penulis: menambah wawasan dan keterampilan penulis dalam menerapkan konsep *CRISP–DM*, teknik *preprocessing* data, dan algoritma *machine learning* untuk menyelesaikan permasalahan klasifikasi di bidang kesehatan.

## 1.5 Sistematika Penulisan

Sistematika penulisan dalam penelitian skripsi ini disusun secara sistematis untuk memberikan alur pembahasan yang logis, terarah, dan mudah dipahami. Struktur penulisan mencakup lima bab utama yang saling berkaitan dan membentuk satu kesatuan analisis ilmiah yang utuh, mulai dari perumusan masalah hingga kesimpulan hasil penelitian. Berikut penjelasan singkat mengenai isi setiap bab:

Bab 1 Pendahuluan, berisi latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

Bab 2 Tinjauan Pustaka, berisi teori-teori pendukung yang relevan dengan penelitian, antara lain konsep dasar diabetes melitus, *data mining*, *machine learning*, *Pima Indians Diabetes Dataset*, algoritma *Artificial Neural Network* dan *Decision Tree*, serta kerangka kerja *CRISP–DM*, serta rangkuman penelitian terdahulu yang berkaitan.

Bab 3 Metodologi Penelitian, berisi penjelasan mengenai rancangan penelitian, sumber dan karakteristik data, tahapan *preprocessing*, penerapan kerangka kerja *CRISP–DM*, serta prosedur pembangunan dan evaluasi model *Artificial Neural Network* dan *Decision Tree*.

Bab 4 Hasil dan Pembahasan, berisi penyajian hasil eksperimen, evaluasi kinerja model, analisis komparatif antara *Artificial Neural Network* dan *Decision Tree*, serta interpretasi hasil penelitian.

Bab 5 Kesimpulan dan Saran, berisi kesimpulan utama yang diperoleh dari penelitian ini serta saran untuk pengembangan penelitian selanjutnya.