BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi terus mengalami kemajuan seiring dengan perubahan zaman. Hal ini dipengaruhi oleh meningkatnya kebutuhan masyarakat terhadap informasi yang cepat dan akurat dalam menunjang aktivitas sehari-hari. Teknologi memiliki peran penting, baik saat ini maupun di masa depan, seperti yang dapat dilihat dari penerapannya di berbagai sektor, salah satunya di bidang kesehatan (S. Id, 2023). Dalam bidang kesehatan, telah mendorong pemanfaatan *machine learning* (ML) sebagai pendekatan berbasis data untuk mendukung proses diagnosis penyakit dan pengambilan keputusan klinis secara lebih cepat dan akurat (Banurea et al., 2023). ML merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem komputer belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dalam bidang kesehatan, ML banyak digunakan untuk menganalisis rekam medis, mengenali pola-pola tersembunyi dalam data klinis, serta memprediksi risiko penyakit secara lebih akurat dan efisien. Salah satu tugas utama dalam ML adalah klasifikasi, yaitu proses pengelompokan data ke dalam kategori tertentu berdasarkan karakteristik atau fitur yang dimilikinya (Magoulas & Prentza, 2015). Dalam klasifikasi penyakit, ML dapat memanfaatkan berbagai jenis data, baik data kuantitatif maupun kategorikal, seperti hasil pemeriksaan laboratorium, karakteristik demografis pasien, serta data klinis lainnya. Oleh karena itu, klasifikasi berbasis ML menjadi pendekatan yang banyak digunakan dalam pengolahan data kesehatan. Berbagai algoritma klasifikasi telah dikembangkan dan digunakan, mulai dari algoritma linear seperti *Logistic*

Regression dan *Support Vector Machine* (SVM), hingga algoritma non-linear dan *ensemble* seperti *Random Forest* dan *XGBoost*. Namun, performa suatu algoritma klasifikasi tidak hanya ditentukan oleh metode yang digunakan, tetapi juga sangat dipengaruhi oleh karateristik *dataset* seperti distribusi kelas, jumlah fitur, serta kompleksitas hubungan antar variabel, Oleh karena itu pemahaman terhadap kondisi dan karakteristik data menjadi langkah awal yang penting sebelum menentukan algoritma yang akan digunakan. Selain itu, penerapan teknik *data handling* yang tepat menjadi tahapan penting sebelum dan selama proses pemodelan. Teknik *data handling* mencakup penanganan *missing value*, penyeimbangan distribusi kelas, transformasi data, standardisasi data, penanganan *outlier*, serta seleksi fitur untuk mengurangi atribut yang tidak relevan (Jarmakovica, 2025).

Dalam konteks ini, tidak terdapat satu algortima yang selalu memberikan hasil terbaik untuk semua permasalahan. Hal ini sejalan dengan No Free Lunch Theorem yang dikemukakan oleh Wolpert dan Macready (1997), yang menyatakan bahwa tidak ada satu algoritma *machine learning* yang selalu memberikan performa terbaik pada seluruh jenis permasalahan. Oleh karena itu, pendekatan komparatif antar algoritma menjadi penting untuk menentukan metode yang paling sesuai dengan *dataset* dan tujuan penelitian tertentu (Singh et al., 2024).

Random Forest dan *XGBoost* merupakan dua algoritma *ensemble* yang banyak digunakan karena kemampuannya dalam menangani data non-linear dan hubungan kompleks antar fitur. Dalam konsep yang diperkenalkan oleh (Breiman, 2001) *Random Forest* merupakan algoritma *ensemble* berbasis *bagging* yang menggabungkan sejumlah decision tree untuk meningkatkan stabilitas model dan

mengurangi risiko overfitting. Sementara itu, dalam teori dasar yang ditemukan oleh (Chen & Guestrin, 2016) *XGBoost* merupakan algoritma *ensemble* berbasis boosting yang membangun model secara bertahap menggunakan optimasi gradien, sehingga lebih sensitif terhadap pengaturan parameter namun berpotensi menghasilkan akurasi yang lebih tinggi.

Algoritma *Random Forest* dan *XGBoost* termasuk sebagai algoritma prediksi yang populer dalam berbagai penelitian klasifikasi, kedua algoritma ini mampu menghasilkan performa klasifikasi yang baik. Beberapa penelitian sebelumnya menunjukkan bahwa performa *Random Forest* dan *XGBoost* dapat bervariasi tergantung pada konteks data dan permasalahan yang diteliti, dimana membandingkan *XGBoost*, *LightGBM*, *Gradient Boosting*, dan *Artificial Neural Network* untuk memprediksi resistansi obat berdasarkan mutasi genomik. Hasil menunjukkan bahwa *XGBoost* unggul dibanding model lain dengan akurasi 97%, recall 97%, spesifisitas 97%, dan F1- score 0,93 (Paredes- Gutierrez et al., 2025). Studi klasifikasi stunting pada anak usia di bawah lima tahun dimana Penelitian ini membandingkan tiga algoritma *Random Forest*, KNN, *XGBoost* pada *dataset* Kaggle anak usia <5 , menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, serta *F1-score*. Hasil akhir *Random Forest* menunjukkan performa terbaik dengan *accuracy* 99,95 %, *precision* 99,89 %, *recall* 99,94 %, serta *F1-score* 99,91 %, mengalahkan KNN serta *XGBoost* (Syahfitri et al., 2024).

Sejalan dengan penelitian-penelitian tersebut, studi oleh (Abdurrahman et al., 2022) pada klasifikasi penyakit diabetes menunjukkan bahwa *XGBoost* tanpa *hyperparameter tuning* menghasilkan akurasi sebesar 75%. Namun, setelah dilakukan *hyperparameter* tuning menggunakan *Grid Search*, akurasi model

meningkat secara signifikan menjadi 95%. Hasil ini memperkuat temuan bahwa optimasi *hyperparameter* merupakan faktor penting dalam meningkatkan performa algoritma *XGBoost*, meskipun efektivitasnya tetap bergantung pada karakteristik *dataset* yang digunakan.

Berdasarkan berbagai hasil penelitian sebelumnya, terlihat bahwa belum terdapat kesimpulan tunggal mengenai algoritma yang paling unggul antara *Random Forest* dan *XGBoost*. Perbedaan pendekatan *ensemble* yang digunakan, yaitu bagging pada *Random Forest* dan boosting pada *XGBoost*, menyebabkan perbedaan karakteristik performa pada masing-masing algoritma. Oleh karena itu, diperlukan penelitian komparatif untuk menganalisis perbedaan performa kedua algoritma tersebut secara lebih objektif, termasuk dengan mempertimbangkan penerapan *hyperparameter* tuning. Di sisi lain, pemilihan kedua algoritma ini didasari oleh efisiensi komputasinya yang baik, sehingga efektif untuk diterapkan pada *dataset* seperti *dataset* kesehatan dalam penelitian ini, tanpa memerlukan sumber daya komputasi yang besar (Grinsztajn & Oyallon, 2022).

Dalam penelitian ini, komparasi *Random Forest* dan *XGBoost* difokuskan pada klasifikasi penyakit berdasarkan data kuantitatif. Sebagai studi kasus digunakan data pasien penyakit Diabetes, merupakan penyakit kelainan metabolisme yang ditandai dengan tingginya kadar gula dalam darah karena kekurangan insulin, atau resistensi insulin. Efek dari penyakit diabetes antara lain adalah kerusakan pada organ tubuh seperti ginjal, jantung, dan saraf (Pengobatan, 2021). Prevalensi diabetes di Indonesia menunjukkan tren peningkatan yang signifikan dari waktu ke waktu. Berdasarkan data Riset Kesehatan Dasar, prevalensi diabetes pada penduduk Indonesia tercatat sebesar 6,9% pada tahun

2013, kemudian meningkat menjadi 10,9% pada tahun 2018. Data terbaru dari *International Diabetes Federation* tahun 2024 mencatat angka prevalensi semakin tinggi, yakni mencapai 11,3%. Hasil Survei Kesehatan Indonesia tahun 2023 yang menunjukkan bahwa prevalensi diabetes telah mencapai 11,7%, namun hanya sekitar 1 dari 4 hingga 5 penderita yang mengetahui status penyakitnya. Tingginya jumlah penderita yang belum terdiagnosis menunjukkan pentingnya deteksi dan klasifikasi penyakit secara dini untuk mencegah komplikasi yang lebih berat (Zamani et al., 2024). *Dataset* yang digunakan berasal dari *Mendeley Data* dengan judul Diabetes *Dataset* yang dipublikasikan pada 18 Juli 2020. *Dataset* ini memiliki 12 variabel medis dan kelas target Diabetes, Non-Diabetes, serta Prediksi Diabetes, sehingga dapat digunakan secara langsung untuk membandingkan kinerja algoritma *Random Forest* dan *XGBoost*. Studi kasus Diabetes dipilih karena karakteristik datanya yang kompleks dan bersifat non-linier, sehingga relevan untuk menguji kemampuan kedua algoritma dalam menangani permasalahan klasifikasi data kesehatan.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk melakukan komparasi performa algoritma *Random Forest* dan *XGBoost* dalam klasifikasi penyakit Diabetes, serta penerapan *Hyperparameter tuning* diharapkan bisa meningkatkan ketepatan dan kekuatan model dalam mengklasifikasi data. Evaluasi performa dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Diharapkan hasil penelitian ini dapat menjadi referensi dalam pemilihan metode klasifikasi yang efektif pada kasus serupa serta mendukung pengembangan sistem deteksi diabetes yang lebih akurat dan efisien di masa mendatang.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, adapun permasalahan yang diidentifikasi dalam penelitian ini adalah:

1. Pemilihan algoritma machine learning yang tepat merupakan aspek penting dalam klasifikasi data kesehatan. Berdasarkan No Free Lunch Theorem (Wolpert & Macready, 1997), tidak ada satu algoritma yang selalu unggul pada semua permasalahan, sehingga diperlukan pendekatan komparatif untuk mengevaluasi kinerja algoritma pada kasus tertentu.
2. Perbedaan pendekatan *ensemble learning* yang digunakan oleh *Random Forest* dan *XGBoost*, yaitu *bagging* dan *boosting*, berpotensi menghasilkan perbedaan performa dalam menangani data kesehatan yang memiliki banyak variabel serta hubungan antar variabel yang bersifat non-linier, seperti pada data penyakit Diabetes. Dengan demikian, diperlukan analisis komparatif untuk menilai kinerja kedua algoritma tersebut secara objektif berdasarkan metrik evaluasi yang relevan.
3. Diabetes penyakit kronis dengan prevalensi yang terus meningkat dan berisiko menimbulkan berbagai komplikasi serius apabila tidak terdeteksi secara dini. Proses klasifikasi penyakit diabetes berbasis data kesehatan menghadapi tantangan tersendiri, seperti banyaknya variabel medis dan hubungan antar variabel yang bersifat non-linier, sehingga diperlukan algoritma klasifikasi yang mampu menangani kompleksitas data tersebut secara efektif.

Berdasarkan latar belakang yang dipaparkan, maka rumusan masalah dala penelitian ini adalah:

1. Bagaimana perbandingan kinerja algoritma *Random Forest* dan *XGBoost*

dalam mengklasifikasikan kasus Diabetes berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-Score*, serta algoritma manakah yang memiliki performa terbaik?

1.2 Tujuan Penelitian

Berdasarkan rumusan masalah yang dibahasa sebelumnya, adapun tujuan dari dilakukannya penelitian ini adalah untuk:

1. Untuk membandingkan kinerja algoritma *XGBoost* dan *Random Forest* dalam klasifikasi Diabetes berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-Score*, serta menentukan algoritma dengan performa terbaik.

1.3 Ruang Lingkup Penelitian

Adapun ruang lingkup penelitian yang diterapakn untuk penelitian ini agar pembahasan lebih berfokus dan spesifikasi unutk mencapai tujuan penelitian, serta tidak meluas dan menyimpang. Batasan-batasan tersebut antara lain:

1. Data yang digunakan dalam penelitian ini berasal dari *Mendeley Data*, yaitu *dataset* diabetes yang diperoleh dari data medis dan hasil pemeriksaan laboratorium pasien. *Dataset* tersebut mencakup informasi klinis dan laboratorium yang relevan dengan kondisi Diabetes.
2. Variabel yang digunakan dalam penelitian ini meliputi atribut medis dan hasil analisis laboratorium, antara lain usia pasien, jenis kelamin, kadar gula darah, indeks massa tubuh (BMI), rasio kreatinin (Cr), urea, kolesterol total,

profil lipid puasa (LDL, VLDL, HDL, trigliserida), serta HbA1c, dengan kelas target berupa Diabetes, Non-Diabetes, dan Prediksi Diabetes.

3. Proses pengolahan data, pelatihan model, dan evaluasi algoritma dilakukan

dengan menggunakan bahasa pemrograman *Python*.

4. Menerapkan *GridSearchCV* untuk *hyperparameter* tuning.

1.4 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian klasifikasi Diabetes menggunakan algoritma *Random Forest* dan *XGBoost* adalah sebagai berikut:

1. Bagi Akademis dan Peneliti Lanjut

Penelitian ini dapat menjadi referensi dalam pengembangan model klasifikasi pada data kesehatan, khususnya dalam melakukan klasifikasi algoritma *machine learning*. Hasil penelitian ini diharapkan dapat memberikan gambaran mengenai perbedaan performa *algoritma Random Forest* dan *XGBoost* dalam menangani data medis yang kompleks dan bersifat non-linear.

2. Bagi Mahasiswa

Meningkatkan pemahaman dalam bidang machine learning, khususnya terkait penerapan algoritma klasifikasi.