BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Diabetes adalah kondisi kronis dengan kadar gula darah tinggi yang berdampak serius pada kesehatan fisik dan mental. Risiko diabetes dipengaruhi oleh kurangnya aktivitas fisik, pola makan tinggi gula dan karbohidrat, serta faktor genetik dan lingkungan. Konsumsi makanan cepat saji, minuman manis, porsi berlebih, kurang tidur, merokok, dan alkohol berlebihan juga meningkatkan risiko. Pencegahan dan pengelolaan diabetes mencakup perubahan gaya hidup seperti pola makan seimbang, olahraga, dan manajemen berat badan. Menurut World Health Organization (WHO), pada tahun 2022, terdapat 537 juta orang dewasa (usia 20-79 tahun) di seluruh dunia yang hidup dengan diabetes, dan jumlah ini diproyeksikan meningkat menjadi 643 juta pada tahun 2030 serta 784 juta pada tahun 2045. Risiko ini semakin meningkat di negara berkembang, di mana prevalensi diabetes diperkirakan akan melonjak dua kali lipat dalam dekade mendatang, dengan lebih dari 80% kematian akibat diabetes terjadi di negara-negara tersebut. Diabetes juga menyebabkan sekitar 6,7 juta kematian pada tahun 2021, dengan 44% penderita tidak terdiagnosis, sehingga meningkatkan risiko komplikasi serius. Berdasarkan data Kementerian Kesehatan dan Atlas IDF edisi ke-10, di Indonesia terdapat 19.465.100 orang dewasa (usia 20-79 tahun) yang menderita diabetes dari total populasi 179.720.500, dengan prevalensi 10,6% atau 1 dari 9 orang [1]. Biaya kesehatan per tahun bagi penderita diabetes di Indonesia hanya 323,8 USD, jauh lebih rendah dibandingkan Australia (5.944 USD) dan Brunei Darussalam (901,3 USD) [1]. Angka kematian terkait diabetes diperkirakan mencapai 236.711 per tahun, sementara 73,7% penderita belum terdiagnosis, meningkatkan risiko komplikasi [1].

Dalam dunia prediksi medis, algoritma Random Forest dan XGBoost sering digunakan karena kemampuan kedua algoritma dalam menganalisis data dengan banyak fitur dan menangani klasifikasi yang kompleks. Pertama, Random Forest dan XGBoost memiliki prinsip kerja yang berbeda, di mana Random Forest menggunakan pendekatan bagging yang membangun banyak pohon keputusan secara paralel, sedangkan XGBoost menggunakan pendekatan boosting yang membangun pohon secara bertahap dengan memperbaiki kesalahan prediksi

sebelumnya [2]. Perbedaan pendekatan ini berpotensi memengaruhi akurasi dan stabilitas prediksi, sehingga penting untuk dibandingkan pada kasus prediksi diabetes yang kompleks. Kedua, kedua algoritma ini dikenal mampu menangani dataset dengan banyak fitur dan data yang tidak seimbang, yang sering ditemui dalam data kesehatan seperti kasus diabetes. Evaluasi terhadap kinerja masing-masing algoritma dalam menangani perbedaan rasio kelas dapat memberikan wawasan tentang algoritma yang lebih optimal. Ketiga, Random Forest dan XGBoost sama-sama populer dalam penelitian medis karena kemampuannya mengidentifikasi fitur-fitur medis seperti pada jurnal yang ditulis oleh Liang Jung Jiang [3] dan Duwi Cahya Putri Buani [4]. Membandingkan keduanya dapat membantu mengidentifikasi algoritma yang tidak hanya lebih akurat, tetapi juga lebih interpretatif untuk mendukung pengambilan keputusan klinis. Penelitian ini bertujuan membandingkan algoritma Random Forest dan XGBoost dalam memprediksi risiko diabetes berdasarkan data medical check-up di Rumah Sakit Pusat Pertamina, serta menawarkan solusi pencegahan yang efektif untuk menurunkan prevalensi diabetes di masa depan.

Dalam penelitian yang dilakukan Maulidah [5], fitur yang digunakan untuk memprediksi penyakit diabetes melitus meliputi variabel seperti jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), Indeks Massa Tubuh (BMI), fungsi keturunan diabetes (DiabetesPedigreeFunction), usia (Age), dan hasil (Outcome). Metode yang diterapkan adalah Support Vector Machine (SVM) dan Naive Bayes, dengan total dataset yang digunakan sebanyak 2000 record yang diambil dari database kesehatan Diabetes Dataset yang dapat diakses melalui Kaggle. Setiap label dalam dataset ini menunjukkan status diabetes, dengan dua kategori yaitu positif dan negatif. Hasil penelitian menunjukkan bahwa akurasi model SVM mencapai 78,04%, sedangkan akurasi model Naive Bayes adalah 76,98%. Namun, informasi mengenai evaluasi lebih lanjut seperti F1-score, recall, dan precision tidak tersedia dalam penelitian ini.

Dalam penelitian yang dilakukan Maulidah [5], fitur yang digunakan untuk memprediksi penyakit diabetes melitus meliputi variabel seperti jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), Indeks Massa Tubuh (BMI), fungsi keturunan diabetes (DiabetesPedigreeFunction), usia (Age), dan hasil (Outcome). Metode yang diterapkan adalah Support Vector Machine (SVM) dan Naive Bayes, dengan total dataset yang digunakan sebanyak 2000 record yang diambil dari

database kesehatan Diabetes Dataset yang dapat diakses melalui Kaggle. Setiap label dalam dataset ini menunjukkan status diabetes, dengan dua kategori yaitu positif dan negatif. Hasil penelitian menunjukkan bahwa akurasi model SVM mencapai 78,04%, sedangkan akurasi model Naive Bayes adalah 76,98%. Namun, informasi mengenai evaluasi lebih lanjut seperti F1-score, recall, dan precision tidak tersedia dalam penelitian ini. Penelitian yang dilakukan oleh Citra Agustina Rahayu, Rudi Hartono, dan Aso Sudiarjo menggunakan dataset dengan 502 data yang diperoleh dari RSUD dr. Soekardjo, dengan 16 atribut seperti umur, glukosa darah puasa, HbA1c, trigliserida, dan kolesterol. Dataset ini terdiri dari 487 data setelah proses pembersihan, dengan 399 kasus diabetes dan 88 kasus non-diabetes [6]. Penelitian ini memanfaatkan algoritma Naive Bayes untuk memprediksi diabetes, menghasilkan akurasi sebesar 95,92%, precision sebesar 97,50%, NPV sebesar 88,89%, recall sebesar 97,50%, dan specificity sebesar 88,89%. Meskipun akurasi yang diperoleh cukup tinggi, penelitian ini terbatas pada analisis menggunakan satu algoritma tanpa penerapan teknik lanjutan seperti hyperparameter tuning, penanganan ketidakseimbangan data, atau evaluasi metrik tambahan seperti F1-score.

Dalam jurnal internasional yang ditulis oleh Liang Jung Jiang[3], model prediksi risiko diabetes dibangun menggunakan fitur-fitur kunci seperti Indeks Massa Tubuh (BMI), usia, tekanan darah sistolik, tekanan darah diastolik, frekuensi olahraga, waktu olahraga, konsumsi makanan pokok, dan status merokok. Metode yang digunakan adalah Random Forest sebagai algoritma klasifikasi utama, bersama dengan algoritma lain seperti eXtreme Gradient Boosting (XGB), K-Nearest Neighbors (KNN), dan Ensemble Learning. Dataset yang digunakan terdiri dari total 252.176 catatan tindak lanjut pasien diabetes yang diperoleh dari sistem manajemen informasi layanan kesehatan masyarakat di Distrik Haizhu, Guangzhou, dari tahun 2016 hingga 2023. Dalam dataset tersebut, terdapat 188.753 label 0 (tanpa diabetes) dan 63.423 label 1 (menderita diabetes). Hasil evaluasi model menunjukkan akurasi sebesar 91,24%, dengan F1-score sebesar 91,72%, recall sebesar 94,22%, dan precision sebesar 89,36%. Model ini menunjukkan kinerja yang tinggi dalam memprediksi risiko diabetes, yang dapat digunakan untuk strategi pencegahan dan pengendalian diabetes di tingkat komunitas.

1.2 Rumusan Masalah

1. Bagaimana performa algoritma Random Forest dan XGBoost dalam mendeteksi diabetes berdasarkan data medical check-up dari Rumah Sakit Pertamina?
2. Apa faktor-faktor yang menyebabkan perbedaan dalam akurasi, presisi, recall, dan f1 score antara algoritma Random Forest dan XGBoost dalam mendeteksi diabetes pada dataset medical check-up dari Rumah Sakit Pertamina?

1.3 Batasan Permasalahan

- Penelitian ini hanya akan menggunakan data medical check-up dari Rumah Sakit Pusat Pertamina sebagai sumber informasi untuk analisis.
- Algoritma yang dianalisis dalam penelitian ini meliputi Random Forest dan XGBoost, tanpa membandingkan algoritma lain.
- Fokus penelitian adalah pada screening diabetes pada individu pasien sesuai data medical check-up dari Rumah Sakit Pusat Pertamina sebagai sumber informasi untuk analisis.
- Penelitian ini tidak akan membahas faktor sosial atau lingkungan yang mempengaruhi prevalensi diabetes, melainkan hanya akan menganalisis data yang tersedia.
- Periode data yang digunakan terbatas pada data medical check-up dari bulan Juli 2023 hingga Agustus 2024.

1.4 Tujuan Penelitian

- Menganalisis performansi algoritma Random Forest dan XGBoost dalam mendeteksi diabetes berdasarkan data medical check-up dari Rumah Sakit Pertamina.
- Mengidentifikasi dan menganalisis faktor-faktor yang menyebabkan perbedaan dalam performa antara algoritma Random Forest dan XGBoost dalam mendeteksi diabetes, seperti arsitektur model, teknik regularisasi,

kecepatan pelatihan, dan pengaruh feature engineering terhadap hasil prediksi.

1.5 Manfaat Penelitian

Secara Teori

- Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan ilmu pengetahuan, khususnya dalam bidang data mining dan machine learning terkait deteksi dini penyakit diabetes. Dengan membandingkan algoritma Random Forest dan XGBoost, penelitian ini akan memperkaya literatur yang ada mengenai efektivitas metode prediksi dalam konteks kesehatan. Selain itu, hasil penelitian ini dapat menjadi acuan bagi penelitian selanjutnya dalam upaya pengembangan algoritma lain untuk diagnosis penyakit.

Secara Praktek

- Penelitian ini akan memberikan informasi yang berguna bagi tenaga medis dan institusi kesehatan dalam meningkatkan proses screening diabetes. Dengan mengetahui algoritma mana yang paling efektif, rumah sakit dapat mengimplementasikan sistem prediksi yang lebih akurat untuk deteksi dini diabetes. Hal ini berpotensi menurunkan angka kejadian diabetes dan komplikasi yang ditimbulkannya, serta meningkatkan kualitas perawatan pasien.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab 1 ini mencakup Latar Belakang Masalah, yang menjelaskan diabetes sebagai kondisi kronis dan data prevalensinya di Indonesia. Rumusan Masalah mengidentifikasi pertanyaan utama terkait optimasi akurasi algoritma Random Forest dan XGBoost dalam screening diabetes. Batasan Permasalahan menetapkan ruang lingkup penelitian, termasuk sumber data

dan fokus analisis. Tujuan Penelitian berfokus pada peningkatan akurasi algoritma dan identifikasi performa algoritma terbaik. Manfaat Penelitian dibagi menjadi manfaat teoritis untuk pengembangan ilmu dan manfaat praktis bagi tenaga medis. Terakhir, Sistematika Penulisan memberikan gambaran umum tentang struktur laporan penelitian.

- Bab 2 LANDASAN TEORI
  Isi Bab 2 ini mencakup landasan teori mengenai diabetes mellitus, serta algoritma Random Forest dan XGBoost. Selain itu, bab ini menjelaskan metode train-test split dalam pengembangan model machine learning, diikuti dengan uji perbandingan akurasi dan kecepatan antara kedua algoritma tersebut dalam memprediksi diabetes.

- Bab 3 METODOLOGI PENELITIAN
  Bab 3 menjelaskan Alur Penelitian melalui flowchart yang menggambarkan tahapan dari pengumpulan data, eksplorasi, persiapan, hingga pembuatan dan evaluasi model. Proses dimulai dengan pengumpulan data, dilanjutkan dengan pemahaman karakteristik data, dan persiapan untuk analisis. Setelah model dibuat, evaluasi dilakukan untuk memastikan kinerja memenuhi standar. Selanjutnya, Alur Preprocessing Data menguraikan langkah-langkah dalam mempersiapkan data, termasuk penggabungan, pemeriksaan data baru, dan penyimpanan data yang sudah diproses. Terakhir, Penjelasan Alur Pembuatan Model menjelaskan proses membangun model, dari persiapan dataset, pemilihan fitur, hingga normalisasi data sebelum menggunakan algoritma Random Forest dan XGBoost.

- Bab 4 HASIL DAN DISKUSI
  Bab 4 berfokus pada Eksperimen dan Hasil Analisis yang dilakukan dalam penelitian. Bagian ini dimulai dengan penjelasan tentang proses pembagian data, termasuk metode pembagian data training dan testing untuk memastikan representasi data yang seimbang. Selanjutnya, tahapan eksperimen dijelaskan, mulai dari penerapan algoritma Random Forest dan XGBoost hingga proses tuning hyperparameter untuk meningkatkan performa model. Bagian ini juga memuat evaluasi hasil eksperimen dengan menggunakan metrik evaluasi seperti akurasi, precision, recall, dan f1-score, baik pada data training maupun testing. Akhirnya, dilakukan analisis hasil untuk membandingkan kinerja kedua algoritma dan memberikan interpretasi

terhadap hasil yang diperoleh.

- Bab 5 KESIMPULAN DAN SARAN

  Bab 5 menyajikan Kesimpulan dan Saran berdasarkan hasil penelitian yang telah dilakukan. Bagian ini dimulai dengan penyampaian kesimpulan utama mengenai performa algoritma Random Forest dan XGBoost dalam memprediksi diabetes, termasuk perbandingan akurasi, f1-score, dan efisiensi masing-masing algoritma. Selanjutnya, diuraikan implikasi dari hasil penelitian ini terhadap pengembangan model prediksi diabetes yang lebih baik di masa depan. Bab ini juga mencakup saran untuk penelitian selanjutnya, seperti eksplorasi dataset yang lebih besar, penggunaan algoritma lain, atau penyempurnaan teknik preprocessing data untuk mendapatkan hasil yang lebih optimal.