

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Diabetes mellitus adalah suatu penyakit atau gangguan metabolisme kronis dengan multi etiologi yang ditandai dengan tingginya kadar gula darah disertai dengan gangguan metabolisme karbohidrat, lipid, dan protein sebagai akibat insufisiensi fungsi insulin. Di Indonesia, diabetes mellitus menjadi penyebab kematian terbanyak keenam, setelah kondisi yang berhubungan dengan persalinan. Pada tahun 2021, sekitar 19,5 juta individu di diagnosa menderita penyakit diabetes mellitus sehingga menjadikan Indonesia peringkat kelima secara global dalam jumlah pasien diabetes tertinggi [1]. Diabetes yang tidak diobati dan tidak teridentifikasi dapat menyebabkan komplikasi serius pada penderitanya [2]. Oleh karena itu, sangat penting untuk melakukan deteksi dini penyakit diabetes mellitus karena jika penyakit ini dibiarkan terlalu lama tanpa pengobatan, dapat mengakibatkan komplikasi berbahaya seperti gagal ginjal, kerusakan fungsi organ lain hingga serangan jantung [3].

Dari perspektif ilmu komputer, proses deteksi dini diabetes mellitus pada manusia dapat dilakukan melalui teknologi informasi dengan menerapkan teknik data *mining*. Data *mining* melibatkan serangkaian tindakan atau proses untuk menemukan hubungan bermakna melalui pola dan tren dalam kumpulan data besar menggunakan berbagai metode dan algoritma [4]. Keunggulan menggunakan sistem informasi dan teknik data *mining* adalah kemampuannya untuk melakukan deteksi dini diabetes mellitus pada manusia dengan cepat. Selain itu, dengan data *mining*, proses deteksi dapat melibatkan beberapa parameter sehingga tidak hanya mengandalkan satu parameter saja. Parameter seperti usia, berat badan, tekanan darah, dan faktor-faktor relevan lainnya dapat dimasukkan, sehingga menghasilkan hasil deteksi yang lebih akurat [5].

Terdapat beberapa penelitian terdahulu yang mengimplementasikan teknik data *mining* dalam prediksi penyakit diabetes mellitus. Penelitian yang dilakukan pada tahun 2023 mengimplementasikan algoritma *Random Forest* dalam deteksi dini penyakit diabetes mellitus pada manusia. Hasil evaluasi kinerja algoritma

Random Forest untuk prediksi penyakit diabetes mendapatkan akurasi sebesar 77,06%, presisi sebesar 71,43%, *recall* sebesar 47,30%, dan kesalahan klasifikasi sebesar 22,94% [6]. Penelitian berikutnya dilakukan pada tahun 2023 mengimplementasikan algoritma *Decision Tree* C4.5 dalam melakukan klasifikasi penyakit diabetes mellitus. Hasil evaluasi kinerja algoritma *Decision Tree* C4.5 untuk klasifikasi penyakit diabetes mendapatkan akurasi sebesar 79%, presisi sebesar 78%, *recall* sebesar 45%, dan *F1-score* sebesar 57% [7]. Kontribusi pada penelitian ini adalah melakukan studi komparasi menggunakan algoritma data *mining* yang telah diimplementasikan pada penelitian sebelumnya yaitu *Random Forest* dan C4.5 dalam memprediksi penyakit diabetes mellitus pada manusia. Sebagai pembeda dari penelitian-penelitian komparasi sebelumnya ditambahkan 1 buah algoritma lagi yaitu *Catboost Classifier* dimana belum ditemukan penelitian sejenis yang melakukan komparasi terhadap algoritma tersebut. Ketiga algoritma tersebut juga dikombinasikan dengan *Adaptive Boosting* sehingga dapat meningkatkan kinerja akurasi dari ketiga algoritma tersebut.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang diuraikan di atas, maka rumusan masalah pada penelitian ini yaitu:

1. Bagaimana kinerja dari algoritma C4.5, *Random Forest*, dan *Catboost Classifier* berbasis *Adaptive Boosting* dalam melakukan prediksi penyakit diabetes mellitus pada manusia?
2. Dari ketiga algoritma data *mining* yang dibahas pada penelitian ini yaitu C4.5, *Random Forest*, dan *Catboost Classifier* berbasis *Adaptive Boosting* manakah yang memiliki kinerja terbaik?

1.3. Tujuan dan Manfaat Penelitian

1.3.1. Tujuan Penelitian

Tujuan yang diharapkan dari penelitian ini yaitu:

1. Untuk mengetahui kinerja dari algoritma C4.5, *Random Forest*, dan *Catboost Classifier* berbasis *Adaptive Boosting* dalam melakukan prediksi penyakit diabetes mellitus pada manusia.

2. Untuk mengetahui dari ketiga algoritma data *mining* yang dibahas pada penelitian ini yaitu C4.5, *Random Forest*, dan *Catboost Classifier* berbasis *Adaptive Boosting* manakah yang memiliki kinerja terbaik.

1.3.2. Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini yaitu:

1. Sebagai kontribusi di bidang kesehatan sehingga dapat menciptakan sistem prediksi yang lebih akurat melalui hasil analisis yang ditemukan pada penelitian ini.
2. Hasil penelitian dapat dijadikan acuan oleh peneliti selanjutnya yang membahas mengenai prediksi penyakit dengan menerapkan algoritma C4.5, *Random Forest*, dan *Catboost Classifier* berbasis *Adaptive Boosting*.

1.4. Batasan Masalah

Berikut ini akan diuraikan batasan masalah pada penelitian ini agar penelitian berjalan sesuai dengan perencanaan yaitu:

1. Dataset pada penelitian ini diambil dari Kaggle Pima Indians Diabetes Database sebanyak 768 data [8].
2. Atribut-atribut *dataset* yang digunakan pada penelitian ini terdiri dari jumlah kehamilan (*pregnant*), kadar gula darah (*glucose*), tekanan darah (*blood pressure*), ketebalan kulit (*skin thickness*), *insulin*, *Body Mass Index* (BMI), indikator riwayat diabetes dalam keluarga (*diabetes pedigree function*), dan umur (*age*).
3. *Output* hasil prediksi terbagi menjadi 2 yaitu 0 (sehat) dan 1 (berpotensi menderita penyakit diabetes mellitus).

1.5. Keterbaruan Penelitian

Berikut ini akan diuraikan keterbaruan dari penelitian ini melalui *literature review* penelitian terdahulu yang berkaitan dengan topik penelitian antara lain:

1. Menurut Taufik Hidayat, Syifa Sabrina Anelia, Rizki Indah Pratiwi, Nadya Salsabila, dan Desta Sandya Prasvita (2021) dalam penelitian yang berjudul “Perbandingan Akurasi Klasifikasi Penyakit Diabetes Menggunakan

Algoritma *Adaboost-Random Forest* dan *Adaboost-Decision Tree* Dengan Imputasi Median dan KNN” mendapatkan hasil evaluasi kinerja algoritma *Adaboost-Random Forest* dengan imputasi *median* mendapatkan akurasi sebesar 77,87%, presisi sebesar 71,47%, *recall* sebesar 100%, dan kesalahan klasifikasi sebesar 22,13% dan algoritma *Adaboost-Random Forest* dengan imputasi KNN mendapatkan akurasi sebesar 77,68%, presisi sebesar 70,77%, *recall* sebesar 100%, dan kesalahan klasifikasi sebesar 22,92%. Sedangkan hasil evaluasi kinerja algoritma *Adaboost-Decision Tree* dengan imputasi *median* mendapatkan akurasi sebesar 79,07%, presisi sebesar 67,11%, *recall* sebesar 100%, dan kesalahan klasifikasi sebesar 25,91% dan algoritma *Adaboost-Random Forest* dengan imputasi KNN mendapatkan akurasi sebesar 79,61%, presisi sebesar 67,82%, *recall* sebesar 100%, dan kesalahan klasifikasi sebesar 25,39%.

2. Menurut P. Suresh Kumar, Anisha Kumari K, Subhashree Mohapatra, Bighnaraj Naik, Janmenjoy Nayak, dan Manohar Misra (2021) dalam penelitian yang berjudul “*CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages*” mendapatkan hasil evaluasi kinerja algoritma *CatBoost Classifier* untuk prediksi penyakit diabetes mendapatkan akurasi sebesar 100%, presisi sebesar 100%, *recall* sebesar 100%, dan *F1-score* sebesar 100%.
3. Menurut Andi, Thamrin, Agus Susanto, Elyzabeth Wijaya, dan Deva Djohan (2023) dalam penelitian yang berjudul “*Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease*” mendapatkan hasil evaluasi kinerja algoritma *Random Forest* untuk prediksi penyakit diabetes mendapatkan akurasi sebesar 77,06%, presisi sebesar 71,43%, *recall* sebesar 47,30%, dan kesalahan klasifikasi sebesar 22,94%.
4. Menurut Nurussakinah dan Faisal (2023) dalam penelitian yang berjudul “Klasifikasi Penyakit Diabetes Menggunakan Algoritma *Decision Tree*” mendapatkan hasil evaluasi kinerja algoritma *Decision Tree* C4.5 untuk klasifikasi penyakit diabetes mendapatkan akurasi sebesar 79%, presisi sebesar 78%, *recall* sebesar 45%, dan *F1-score* sebesar 57%.

