BAB I
PENDAHULUAN

1.1. Latar Belakang

Diabetes telah menjadi salah satu tantangan kesehatan global yang paling mendesak, dengan prevalensi dan insidensinya yang terus meningkat secara mengkhawatirkan. Menurut World Health Organization (2024), diabetes adalah penyakit kronis yang terjadi ketika pankreas tidak menghasilkan cukup insulin atau ketika tubuh tidak dapat menggunakan insulin secara efektif. Permasalahan ini semakin kompleks dengan adanya kesulitan dalam penanganan dan pendataan warga yang berpotensi menderita diabetes, terutama di wilayah terpencil, fasilitas kesehatan dasar, dan dalam kegiatan survei kesehatan masyarakat secara langsung. Secara umum, terdapat dua jenis diabetes yaitu diabetes mellitus dan diabetes insipidus. Diabetes insipidus (DI) merupakan gangguan keseimbangan air yang dapat bersifat kongenital atau didapat, dengan gejala utama berupa rasa haus berlebihan, poliuria, kelelahan, dehidrasi, dan penurunan berat badan (Jasmeen et al., 2024). Sementara itu, diabetes mellitus yang menjadi fokus utama penelitian ini ditandai dengan kadar glukosa darah yang tinggi dan muncul dalam berbagai bentuk dengan karakteristik penyebab, perkembangan, dan strategi manajemen yang berbeda (Mukamana, 2024). Diabetes mellitus diklasifikasikan menjadi diabetes tipe 1 (T1D), diabetes tipe 2 (T2D), dan diabetes gestational (Okechukwu et al., 2023).

Situasi diabetes di Indonesia menunjukkan kondisi yang sangat mengkhawatirkan dan memerlukan perhatian serius dari berbagai pihak. Berdasarkan data International Diabetes Federation (IDF), Indonesia menempati peringkat kelima negara dengan jumlah penderita diabetes terbanyak di dunia, dengan 19,5 juta penderita pada tahun 2021, dan diproyeksikan akan meningkat drastis menjadi 28,6 juta pada tahun 2045 jika tidak dilakukan intervensi yang tepat (Sehat Negeriku Kemkes, 2024; Harbuwono, 2024). Temuan ini diperkuat oleh penelitian yang dilakukan oleh (Wahidin et al., 2024) menunjukkan proyeksi prevalensi diabetes di Indonesia akan meningkat secara signifikan dari 9,19% pada tahun 2020 (18,69 juta kasus) menjadi 16,09% pada tahun 2045 (40,7 juta kasus), yang menggambarkan

peningkatan sebesar 75,1% selama 25 tahun dengan rata-rata peningkatan 3% per tahun.

Kompleksitas permasalahan diabetes terletak pada potensi komplikasi jangka panjang yang dapat timbul apabila tanda-tanda dan gejala awal diabaikan. Komplikasi tersebut meliputi masalah ginjal, gangguan penglihatan, dan penyakit kardiovaskular yang dapat mempengaruhi jantung dan pembuluh darah (Addo et al., 2024). Manifestasi klinis diabetes pada tahap awal umumnya berupa poliuria, polidipsia, penurunan berat badan mendadak, polifagia, penglihatan kabur, dan *delayed wound healing*. Namun, gejala-gejala tersebut seringkali bersifat non-spesifik dan dapat menyerupai berbagai kondisi medis lainnya, sehingga menyulitkan proses diagnosis dini. Kondisi inilah yang menjadikan pendekatan machine learning sebagai solusi yang sangat relevan dan strategis dalam konteks deteksi dini diabetes.

Penggunaan *machine learning* dalam memprediksi dan menentukan status atau target apakah seseorang diabetes atau tidak berdasarkan gejala yang serupa menjadi pilihan yang lebih unggul dibandingkan dengan pendekatan tradisional. Pendekatan konvensional sering kali bergantung pada metode statistik yang tidak dapat menangkap kompleksitas dan interaksi antara variabel yang ada, seperti menggunakan ukuran rata-rata, median, dan deviasi standar untuk menggambarkan data. Maka dari itu, m*achine learning* menawarkan keunggulan dalam hal akurasi dan kecepatan dalam memproses dan menganalisis data dalam jumlah yang banyak atau besar, serta kemampuan untuk mendeteksi pola yang kompleks dalam sebuah dataset.

Machine learning (ML) adalah bagian dari *artificial intelligence* (AI) yang menggunakan algoritma untuk menganalisis kumpulan data dan membuat model pembelajaran mandiri yang mampu memprediksi hasil dan mengkategorikan informasi secara otomatis. Pemilihan ML dibandingkan *deep learning* (DL) dikarenakan penggunaan data yang lebih sedikit pada penelitian ini, yang apabila menggunakan model DL akan mendapatkan kinerja yang lebih rendah jika dibandingkan dengan algoritma ML (Reber et al., 2023). Sudah banyak pendekatan algoritma *machine learning* pada prediksi penyakit seperti kanker paru, kanker payudara, obesitas, diabetes, stroke, dan penyakit jantung. Penerapan *machine*

learning dalam prediksi penyakit membantu dalam deteksi dini dan intervensi yang tepat waktu, yang pada akhirnya dapat meningkatkan hasil kesehatan pasien. Metode algoritma ML yang umum digunakan dalam deteksi dini diabetes antara lain, algoritma *Random Forest* (RF), *K-nearest neighbor* (KNN), *Support Vector Machine* (SVM), dan *Logistic Regression* (LR) (Addo et al., 2024; Apriliah et al., 2021; Najla Salsabila et al., 2024; Safitri & Praba, 2024).

Namun, meskipun banyak algoritma yang telah digunakan, beberapa kelemahan tetap ada. Misalnya, algoritma seperti *Random Forest* dan KNN yang dapat mengalami *overfitting,* atau kondisi dimana sebuah *machine learning* terlalu fokus pada data pelatihan dan tidak mempelajari detail-detail kecil yang mungkin tidak relevan, yang berakibat model tersebut bekerja sangat baik pada data pelatihan, namun tidak dapat bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya, terutama ketika berhadapan dengan data yang memiliki banyak fitur atau variabel. Selain itu, algoritma tersebut seringkali memerlukan *pre-processing* yang ekstensif, seperti *one-hot encoding* untuk data kategorikal, yang dapat menambah kompleksitas dan waktu pemrosesan.

Categorical Boosting atau CatBoost adalah algoritma pembelajaran mesin berbasis pohon keputusan yang menggunakan teknik boosting (Hancock & Khoshgoftaar, 2020). Algoritma ini dikembangkan oleh Yandex dan dirancang untuk menangani data kategorikal dengan lebih efisien (Yandex, n.d.). Salah satu keunggulan matematis dari CatBoost adalah kemampuan untuk mengatasi *overfitting* melalui penggunaan skema pemrosesan data yang unik, seperti pemrosesan data kategorikal dan penggunaan teknik boosting berbasis gradien yang dioptimalkan dan juga CatBoost dapat mengimplementasikan metode pemrosesan data yang mengurangi bias prediksi, yang sering kali menjadi tantangan bagi algoritma boosting lainnya. Pemilihan algoritma CatBoost untuk penelitian ini didasarkan pada beberapa alasan teknis dan matematis yang membedakannya dari algoritma boosting lainnya, yaitu CatBoost secara khusus dirancang untuk menangani data kategorikal tanpa memerlukan *pre-processing* yang ekstensif seperti *one-hot encoding*, seperti yang diperlukan algoritma lain. Selain itu pengurangan risiko *overfitting* dan bias prediksi yang merupakan masalah umum dalam algoritma boosting lainnya (Vegetti & Meroi, 2024).

Dengan demikian, fokus dari penelitian ini adalah mengetahui pola tertentu dalam gejala-gejala yang dialami oleh pasien diabetes. Selanjutnya, dilakukan perbandingan evaluasi performa dari algoritma *Random Forest,* KNN, XGBoost, AdaBoost, LightGBM, dan *Logistic Regression* dengan algoritma yang diusulkan yaitu *Categorical Boosting* (CatBoost) dikarenakan keenam algoritma tersebut memiliki evaluasi performa yang tinggi saat digunakan untuk prediksi penyakit diabetes. Pemilihan CatBoost juga didasarkan pada keunggulan teknis dan matematisnya yang diharapkan dapat memberikan hasil yang diharapkan dalam penelitian ini. Selain itu, motivasi dilakukannya penelitian ini juga bertujuan untuk membantu penanganan dan pendataan warga-warga yang berpotensi menderita diabetes melalui sistem prediksi berbasis machine learning yang dapat digunakan di berbagai tempat, khususnya wilayah terpencil, fasilitas kesehatan dasar, maupun dalam kegiatan survei kesehatan masyarakat secara langsung. Dengan sistem prediksi yang fleksibel dan portabel ini, diharapkan upaya deteksi dini dan intervensi medis dapat dilakukan lebih cepat, efisien, dan menyeluruh.

1.2. Perumusan Masalah

Berdasarkan uraian yang telah dijelaskan, maka identifikasi masalah dalam penyusunan Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana mengidentifikasi gejala-gejala awal yang paling berpengaruh dalam penyakit diabetes?
2. Bagaimana mengevaluasi efektivitas algoritma CatBoost dalam memprediksi diabetes berdasarkan hasil evaluasi performa yang tinggi?
3. Bagaimana mengembangkan aplikasi berbasis website yang dapat membantu dalam melakukan prediksi awal terhadap risiko diabetes?

1.3. Tujuan

Berdasarkan identifikasi masalah yang telah dijelaskan, tujuan penyusunan Tugas Akhir ini adalah sebagai berikut:

1. Mengidentifikasi gejala-gejala awal yang paling berpengaruh dalam penyakit diabetes
2. Mengevaluasi efektivitas algoritma CatBoost dalam memprediksi dengan hasil evaluasi performa yang tinggi.

3. Mengembangkan aplikasi berbasis website untuk membantu dalam melakukan prediksi awal terhadap risiko diabetes

1.4. Batasan dan Asumsi Penelitian

Batasan pada penelitian ini mencakup waktu penyusunan penelitian yang dimulai dari bulan September hingga bulan Juli. Dengan dataset primer yang berhasil kumpulkan dari bulan Febuari hingga bulan Maret.

Kemudian, diasumsikan bahwa semua pasien yang menjadi responden telah memberikan informasi yang lengkap dan benar mengenai gejala yang dialami.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat berupa, wawasan dan pengetahuan mengenai penerapan algoritma ML dalam mendeteksi diabetes berdasarkan gejala yang ada. Penelitian ini juga dapat menjadi referensi bagi penelitian-penelitian selanjutnya yang ingin mengembangkan metode deteksi dini penyakit menggunakan teknologi ML.

Penelitian ini juga diharapkan dapat membantu para pengguna dalam melakukan deteksi dini diabetes dengan cepat, sehingga dapat memberikan intervensi yang tepat waktu dan mengurangi risiko komplikasi jangka panjang.

1.6. Sistematika Penulisan

Sistematika penulisan Tugas Akhir ini terdiri dari enam bab, dengan uraian sebagai berikut:

- Bab I Pendahuluan

  Berisi latar belakang masalah, rumusan masalah, tujuan penelitian, batasan dan asumsi penelitian, manfaat penelitian, serta sistematika penulisan dari laporan tugas akhir ini.
- Bab II Landasan Teori

  Membahas teori-teori yang relevan dengan topik penelitian, seperti konsep dasar diabetes, algoritma machine learning yang digunakan, perbandingan algoritma, metrik evaluasi, serta penjelasan metode Black Box Testing sebagai pendekatan pengujian sistem.
- Bab III Alur Pemodelan

Menjelaskan metode penelitian secara menyeluruh, mulai dari studi literatur, pengumpulan dan pemrosesan data, skenario pembagian data, pengembangan sistem berbasis website, serta proses evaluasi sistem menggunakan metode pengujian yang telah dirancang.

- Bab IV Pengumpulan dan Pengolahan Data

  Menguraikan proses akuisisi data primer dan sekunder, normalisasi dan transformasi data, seleksi fitur menggunakan Chi-Square, serta pelatihan model awal sebagai baseline evaluasi performa.
- Bab V Analisis dan Pembahasan

  Berisi analisis hasil penelitian, verifikasi dan validasi sistem, evaluasi performa model, perbandingan hasil antara model CatBoost dengan algoritma lain, identifikasi gejala yang paling berpengaruh, serta pengujian website menggunakan metode Black Box Testing.
- Bab VI Kesimpulan dan Saran

  Menyajikan ringkasan dari hasil penelitian serta saran-saran untuk pengembangan sistem lebih lanjut di masa depan