BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes adalah penyakit yang disebabkan karena terlalu banyak kadar gula dalam darah. Gula yang seharusnya bisa diolah menjadi sumber energi karena suatu hal tidak dapat diserap dalam tubuh, sehingga menyebabkan penumpukan dalam sel tubuh. Menurut *International Diabetes Federation* (IDF), pada tahun 2021 total penderita diabetes di dunia adalah 537 juta dan Indonesia sendiri ada 19 juta serta akan terus bertambah setiap tahunnya[1][2].

Penyakit diabetes dibagi menjadi 2 jenis, yaitu diabetes tipe 1 dan tipe 2. Diabetes tipe 1 terjadi saat tubuh tidak bisa memproduksi insulin yang dibutuhkan tubuh, tipe 1 ini biasanya ditemukan pada remaja dan bisa dibantu dengan terapi insulin[3]. Diabetes tipe 2 terjadi karena tubuh tidak bisa memproses insulin yang ada sehingga gula tidak dapat diproses dan tetap berada dalam darah. Penderita diabetes tipe 2 juga memiliki resiko yang lebih tinggi terkena kanker[4]. Diabetes bisa diantisipasi dengan makan makanan sehat, sering berolahraga dan menjaga berat badan[5].

Penyebab diabetes sendiri ada beberapa faktor, antara lain genetik, lingkungan, gaya hidup, obesitas dan masih banyak lagi[6]. Gejala yang sering dialami penderita diabetes adalah sering buang air kecil (*poliuri*), cepat merasa haus, pandangan kabur, selalu lelah dan berat badan menurun[5]. Seiring berjalannya waktu, diabetes bisa merusak pembuluh darah yang ada di saraf, hati, ginjal dan mata. Diabetes juga merupakan penyebab utama dari kebutaan, gagal ginjal, serangan jantung dan stroke[5]. Untuk sekarang ini, belum ada obat dari penyakit diabetes[7]. Yang bisa kita lakukan hanyalah mencegah dan memperlambatnya saja[7]. Walaupun begitu, sekarang kita sudah bisa memprediksi resiko diabetes menggunakan teknologi *Machine Learning*.

Machine Learning adalah cabang ilmu dalam *Artificial Intelligence* yang berfokus untuk meniru bagaimana cara manusia belajar hanya menggunakan data

dan algoritma yang ada[8]. *Machine Learning* sendiri sudah sering dimanfaatkan dalam bidang kesehatan, seperti untuk mendeteksi penyakit[9], mempercepat pemulihan, meringankan stres pada pasien dan sebagainya[10]. Dalam kasus diabetes sendiri, *Machine Learning* sudah beberapa kali digunakan dengan tujuan memprediksi resiko diabetes pada seseorang. Salah satu contohnya jurnal dengan judul "Classification of Diabetes using Machine Learning" yang bertujuan untuk mengetes akurasi dari algoritma *Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), Support Vector Machine (Sigmoid Kernel), Gradient Boosting, KKN (K Means Clustering)* dan *Linear Discriminant Analysis*. Hasilnya algoritma *Naïve Bayes* memiliki tingkat keakuratan paling tinggi dengan angka 79,87%[7]. *Dataset* yang dipakai pada penelitian diatas adalah Pima Indians. Kemudian dalam jurnal tahun 2023 "Comparison of Machine Learning Algorithms and Feature Visualization Analysis for Diabetes Risk Prediction", parameter evaluasi yang digunakan yaitu *Accuracy, Recall* dan *AUC* (*Area Under the Curve*). Penelitian ini menggunakan Diabetes Health Indicators *dataset* dan algoritma *Decision Tree, Random Forest, XGBoost, Logistic, SVM (Support Vector Machine)* & *DNN (Deep Neural Networks)*. Dari tes yang dilakukan, *DNN* memiliki keunggulan dibanding algoritma yang lain dengan akurasi 75,49% dan *AUC* yang mencapai 75,20%[11].

Dari penelitian yang sudah pernah dilakukan, dapat dikatakan bahwa hasil dari tiap tes tergantung dari *dataset* dan algoritma yang digunakan serta bagaimana *preprocessing* data yang dilakukan. Peneliti berencana melakukan penelitian untuk menganalisis keefektifan penggunaan metode *preprocessing* data menggunakan *Label Encoding* & *One Hot Encoding* pada algoritma algoritma *Support Vector Machine*, *Random Forest*, dan *Naïve Bayes*. Kemudian akan ada beberapa parameter yang akan digunakan untuk mengukur kinerjanya dengan menggunakan nilai *Accuracy, Precision, Recall, F-1 Score* dan *Confusion Matrix*. *Dataset* yang akan digunakan merupakan *dataset* klasifikasi penyakit diabetes yang diambil dari Kaggle.

Penelitian ini diharapkan mampu memberikan pemahaman mendalam tentang kinerja algoritma ketika menggunakan *preprocessing* data dengan metode *One Hot Encoding* dan *Label Encoding* untuk deteksi dini diabetes. Oleh sebab itu peneliti akan melakukan penelitian dengan judul "Analisis Kinerja Algoritma *Machine Learning* dengan Metode *Preprocessing* Data *One Hot Encoding* dan *Label Encoding* dalam Prediksi Penyakit Diabetes "

1.2 Rumusan Masalah

Bagaimana hasil kinerja algoritma *machine learning* yaitu, *Support Vector Machine* (*SVM*), *Random Forest* dan *Naive Bayes* dengan metode *preprocessing* data menggunakan *One Hot Encoding* dan *Label Encoding* pada kasus klasifikasi penyakit diabetes ?

1.3 Batasan Masalah

1. Algoritma yang dipakai dalam penelitian ini adalah *Support Vector Machine, Random Forest* dan *Naive Bayes* serta menggunakan metode *preprocessing* data *Label Encoding* dan *One Hot Encoding*.
2. Tidak mengubah fungsi parameter algoritma yang digunakan atau dengan kata lain menggunakan *default* fungsi algoritma yang dipakai.
3. Parameter evaluasi model yang digunakan adalah *accuracy, precision, recall, f-1 score* dan *confusion matrix*.
4. *Dataset* yang akan dipakai adalah *dataset* "Diabetes Prediction Dataset" yang diunggah oleh Mohammed Mustafa pada platform Kaggle yang memiliki 8 atribut yaitu "age", "gender", "bmi", "hypertension", "heart disease", "HbA1c level", "blood glucose level" dan 1 label yaitu "diabetes".

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk melakukan perbandingan kinerja antara tiga algoritma *machine learning*, yaitu *Support Vector Machine (SVM), Random Forest*, dan *Naive Bayes*, dalam konteks penggunaan dua metode *encoding* yang berbeda,

yakni *One Hot Encoding* dan *Label Encoding* pada kasus klasifikasi penyakit diabetes.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan pemahaman mendalam tentang tingkat efisiensi dari algoritma yang diuji serta memungkinkan untuk menemukan korelasi *preprocessing* data menggunakan *Label Encoding* dan *One Hot Encoding* dengan hasil performa algoritma untuk deteksi dini diabetes. Penelitian ini juga diharapkan bisa menjadi referensi bagi orang lain, terkhusus di bidang analisis kinerja algoritma.

1.6 Sistematika Penulisan

Untuk mempermudah dalam memahami skripsi ini, maka penulis materi disusun dengan sistematika sebagai berikut:

BAB I PENDAHULUAN, berisi Latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi tinjauan pustaka, dasar-dasar teori yang digunakan dan dijadikan dasar penelitian dalam skripsi ini.

BAB III METODE PENELITIAN, didalamnya terdapat tinjauan umum tentang alur penelitian, analisis masalah, tahap rancangan, serta alat dan bahan yang digunakan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN, bab ini merupakan tahapan yang penulis lakukan dalam hasil penelitian yang dicapai, serta menjelaskan hasil uji coba rancangan yang telah dibuat.

BAB V PENUTUP, berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian yang penulis berikan untuk peneliti selanjutnya.