GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
doi

PENDAHULUAN

Diabetes Mellitus (DM) merupakan penyakit kronis dengan kadar gula darah melebihi batas normal yang dapat menimbulkan komplikasi serius apabila tidak ditangani [1][2]. Menurut *International Diabetes Federation (IDF)* melaporkan bahwa pada tahun 2024 terdapat sekitar 589 juta penderita diabetes di dunia dan diperkirakan meningkat menjadi 853 juta pada tahun 2050. Pada tahun yang sama lebih dari 3,4 juta kematian akibat komplikasi diabetes dan 43% penderita diabetes belum terdiagnosis [3]. Kondisi ini menegaskan urgensi deteksi dini sebagai upaya pencegahan serta menekan risiko komplikasi dan kematian. Diabetes umumnya disertai dengan gejala seperti polyuria, polydipsia, polyphagia, penurunan berat badan, kelelahan, pandangan kabur, dan luka yang sulit sembuh [4][5]. Gejala-gejala tersebut menjadi indikator penting dalam proses diagnosis diabetes berbasis gejala, yang dapat dimanfaatkan untuk mendukung sistem deteksi dini penyakit diabetes.

Seiring berkembangnya teknologi, proses deteksi dini penyakit diabetes dapat dibantu dengan metode berbasis data. Salah satu pendekatan yang banyak dipakai adalah *data mining*. *Data mining* adalah tahapan mengekstraksi dan menganalisis informasi relevan dari basis data besar dengan pendekatan statistik, matematika, *machine learning*, dan kecerdasan buatan [6]. *Data mining* menggunakan berbagai pendekatan untuk mengumpulkan informasi, termasuk deskripsi, estimasi, pengelompokan, asosiasi, prediksi, dan klasifikasi [7]. Pada berbagai pendekatan yang digunakan dalam *data mining*, klasifikasi menjadi salah satu metode yang kerap dimanfaatkan pada berbagai bidang, salah satunya di bidang kesehatan karena mampu mengelompokkan objek ke dalam kelompok tertentu, misalnya positif atau negatif suatu penyakit, berdasarkan informasi data yang dimiliki. Dengan demikian, penggunaan metode klasifikasi dapat membantu upaya dalam deteksi dini penyakit diabetes.

Klasifikasi adalah proses mengkategorikan objek-objek berdasarkan ciri-ciri yang sama dan mengembangkan model maupun fungsi yang mampu mendefinisikan, membedakan, dan memperkirakan kelas objek yang sebelumnya tidak diketahui [8][9]. Terdapat beberapa metode klasifikasi yaitu Naïve Bayes, *K-Nearest Neighbor (KNN), Decision Tree (C4.5, C5.0), Artificial Neural Network (ANN), Random Forest,* dan *Support Vector Machine (SVM)* [10]. Penelitian ini menerapkan metode Naïve Bayes, metode ini dipilih karena kecepatan kalkulasi, kesederhanaan algoritma, dan kemampuannya mencapai *accuracy* tinggi [11]. Klasifikasi Naïve Bayes merupakan metode statistik untuk mengidentifikasi keanggotaan suatu kelas. Prinsip klasifikasi ini diperkenalkan oleh Thomas Bayes, seorang ilmuwan asal Inggris, yaitu memperkirakan probabilitas masa depan berlandaskan peristiwa sebelumnya yang kemudian dikenal luas dengan Teorema Bayes [12]. Beberapa penelitian sebelumnya menunjukkan metode Naïve Bayes mampu menghasilkan *accuracy* model klasifikasi dengan baik. Pada penelitian oleh [13] menggunakan metode Naïve Bayes untuk mengklasifikasi penyakit diabetes pada perempuan, memperoleh *accuracy* 78,50%. Penelitian lain oleh [14] mengevaluasi prediksi risiko diabetes tahap awal menggunakan metode Naïve Bayes dengan *cross-validation* mencapai tingkat *accuracy* sebesar 87,88%. Meskipun metode Naïve Bayes telah terbukti memiliki keunggulan dalam klasifikasi penyakit diabetes, penelitian lebih lanjut diperlukan untuk mengevaluasi efektivitasnya pada berbagai kondisi dataset, termasuk yang telah melalui seleksi variabel. Tahapan ini penting untuk mengidentifikasi variabel yang berpengaruh dalam dataset, sehingga model klasifikasi fokus pada variabel yang memiliki pengaruh signifikan.

Seleksi variabel merupakan proses identifikasi dan seleksi subset variabel bebas yang berpengaruh dari kumpulan data besar yang akan diterapkan untuk membentuk model klasifikasi [15]. Seleksi variabel bertujuan mengurangi variabel yang tidak relevan, mempercepat proses pelatihan model, dan meningkatkan kinerja pengklasifikasi [16]. Akan tetapi, menemukan variabel yang optimal bukanlah perkara yang mudah karena berpotensi menimbulkan tingginya bias atau varians [17]. Oleh karena itu, dibutuhkan strategi seleksi variabel yang tepat dan relevan dengan data. Chi-Square dan *Mutual Information* adalah dua pendekatan pemilihan variabel berbasis filter yang banyak digunakan [18]. Metode Chi-Square adalah teknik seleksi variabel yang digunakan untuk menilai seberapa besar hubungan antara sebuah variabel dengan kelas [19]. Menurut [20] pada penelitiannya membuktikan

Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
doi

bahwa penerapan seleksi variabel Chi-Square pada Naïve Bayes meningkatkan *accuracy* mencapai 2,38%. Penelitian serupa oleh [21] menunjukkan adanya peningkatan *accuracy* sebesar 6,78%.

Hasil penelitian sebelumnya menegaskan bahwa penggunaan metode Naïve Bayes dengan seleksi variabel Chi-Square mampu meningkatkan performa *accuracy* model klasifikasi. Berdasarkan uraian tersebut, permasalahan dalam penelitian ini adalah belum adanya kajian yang secara spesifik menganalisis pengaruh perbedaan taraf signifikansi pada seleksi variabel Chi-Square terhadap performa *accuracy* metode Naïve Bayes dalam klasifikasi penyakit diabetes. Berdasarkan permasalahan tersebut, penelitian ini melakukan klasifikasi penyakit diabetes menggunakan metode Naïve Bayes dengan dan tanpa seleksi variabel Chi-Square pada tiga tingkat signifikansi, yaitu $\alpha = 0{,}05$, $\alpha = 0{,}01$, dan $\alpha = 0{,}001$. Penelitian ini bertujuan membandingkan performa *accuracy* Naïve Bayes tanpa seleksi variabel dengan Naïve Bayes yang menggunakan seleksi variabel Chi-Square pada berbagai tingkat signifikansi, sehingga dapat diketahui sejauh mana seleksi variabel berpengaruh terhadap peningkatan *accuracy* model klasifikasi penyakit diabetes.