1. Pendahuluan

1.1 Latar Belakang

Diabetes adalah penyakit berbahaya yang ditandai dengan peningkatan kadar gula dalam darah atau glukosa darah tinggi. Diabetes merupakan suatu kondisi di mana pankreas tidak dapat menghasilkan jumlah insulin yang dibutuhkan untuk mengatur jumlah gula dalam darah. Insulin merupakan hormon yang diproduksi oleh pankreas, yang berfungsi untuk mengubah kadar glukosa berlebih dari aliran darah menjadi energi. Diabetes dapat disebabkan oleh pola hidup yang tidak sehat. Pola hidup yang mengonsumsi makanan cepat saji, makanan berkarbohidrat tinggi, gula berlebih, minuman bersoda, minuman beralkohol, dan kurang berolahraga. Diabetes dapat menyebabkan komplikasi pada bagian tubuh yang dapat mengakibatkan kematian. Penyakit jantung, gagal ginjal, kehilangan penglihatan dan kerusakan saraf termasuk kemungkinan komplikasi dari penyakit diabetes. Orang dewasa dengan penyakit diabetes memiliki peningkatan risiko serangan jantung dan stroke dua hingga tiga kali lipat. Pada wanita hamil, diabetes yang tidak terkontrol meningkatkan risiko kematian janin dan komplikasi lainnya [1].

Pada tahun 2014, Jumlah penderita diabetes meningkat dari 108 juta pada tahun 1980 menjadi 422 juta. Pada tahun 2016, diperkirakan 1,6 juta kematian secara langsung disebabkan oleh diabetes [2]. Pada tahun 2019, sekitar 463 juta orang dewasa yang berusia 20-79 tahun hidup dengan diabetes, diperkirakan akan meningkat menjadi 700 juta orang pada tahun 2045. 1 dari 2 (232 juta) orang dengan diabetes tidak terdiagnosis. Lebih dari 20 juta kelahiran hidup (1 dari 6 kelahiran) dipengaruhi oleh diabetes selama kehamilan. 3 dari 4 individu (79%) penderita diabetes tinggal di negara berpenghasilan rendah dan menengah [3].

Pada tahun 2018, dilakukan penelitian oleh Sisodia D. dan Sisodia D. S. dengan judul "Prediction of Diabetes using Classification Algorithms". Penelitian tersebut merancang model yang dapat memprediksi kemungkinan terjadinya diabetes pada penderita diabetes dengan membandingkan tiga metode klasifikasi, yaitu Support Vector Machine (SVM), Naïve Bayes, dan Decision Tree. Hasil dari penelitian menunjukan bahwa metode naïve bayes memiliki tingkat akurasi yang lebih tinggi daripada metode yang lain dengan akurasi 76.30% [4].

Berdasarkan data tersebut, diabetes merupakan salah satu penyakit yang mengancam kesehatan setiap individu di dunia, baik orang dewasa maupun bayi yang baru lahir. Dengan teknologi yang berkembang pesat, banyak metode pembelajaran mesin yang telah sukses untuk mendeteksi penyakit diabetes. Selain permasalahan tersebut, permasalahan di dalam dataset juga dapat mempengaruhi kualitas performa, *missing value* merupakan salah satu faktor yang dapat menurunkan performa [5]. Oleh karena itu, penelitian ini ingin membuat sebuah sistem klasifikasi penyakit diabetes menggunakan metode naïve bayes. Metode naïve bayes dianggap sebagai algoritma yang efektif yang digunakan untuk tujuan klasifikasi [4]. Selain itu, dilakukan perbandingan terhadap teknik penanganan *missing value* pada dataset diabetes.

Penelitian ini bertujuan membuat sistem klasifikasi penyakit diabetes menggunakan metode naïve bayes dan membandingkan performa klasifikasi dengan penggunaan teknik penanganan *missing value* yang berbeda untuk mengetahui performa metode naïve bayes dan teknik penanganan *missing value*.

1.2 Topik dan Batasannya

Topik yang akan dibahas pada penelitian ini adalah mengetahui dan membandingkan pengaruh penggunaan teknik penanganan *missing value*, yaitu dengan teknik menghapus record yang berisi atribut *missing value* dan teknik *unsupervised imputation* yang dilakukan pada saat sebelum dan sesudah partisi data dengan menggunakan metode klasifikasi Naïve Bayes. Data yang digunakan merupakan Dataset Gula Karya Medika yang berjumlah 470 record.

1.3 Tujuan

Tujuan penelitian ini adalah untuk membuat model klasifikasi naïve bayes dan membandingkan hasil performa lima teknik penanganan *missing value* yang dilakukan sebelum dan sesudah partisi data bertujuan mendapatkan hasil performa terbaik.

1.4 Organisasi Tulisan

Bab dua akan menjelaskan studi terkait yang berisikan teori pendukung penelitian. Bab 3 menjelaskan rancangan sistem penelitian. Bab empat akan menjelaskan hasil analisis dan pengujian penelitian. Bab 5 akan menjelaskan kesimpulan yang diperoleh dari proses penelitian beserta saran untuk penelitian selanjutnya.

2. Studi Terkait

2.1 Penelitian Terkait

Penelitian tentang klasifikasi pada penyakit diabetes banyak dilakukan karena identifikasi penderita diabetes yang memerlukan pemeriksaan dokter dengan biaya yang mahal dan waktu yang lama. Dengan adanya penelitian

yang merancang sistem klasifikasi, biaya dan waktu akan berkurang jika penderita diabetes dapat memprediksi diabetes secara cepat. Penelitian tersebut telah banyak dilakukan oleh penelitian sebelumnya dengan menggunakan dataset dan metode klasifikasi yang berbeda, *seperti Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), Naïve Bayes*, dan gabungan beberapa metode.

Penelitian klasifikasi penyakit diabetes pernah dilakukan oleh [6] menggunakan metode *Support Vector Machine* (SVM) dengan dataset *Pima Indians Diabetes Database* (PIDD) dimana hasil klasifikasi menghasilkan akurasi 78%. Pada penelitian [4] menggunakan beberapa metode klasifikasi, seperti *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Decision Tree* untuk mengklasifikasikan penyakit diabetes dan memprediksi hasil dari metode yang digunakan. Hasil yang didapat setelah percobaan, metode *Naïve Bayes* memiliki tingkat akurasi yang lebih tinggi daripada *Support Vector Machine dan Decision Tree* dengan nilai akurasi 76.3%.

Pada penelitian [7] merancang sistem klasifikasi untuk meningkatkan akurasi dengan mengadaptasi klasifikasi berbasis *Gaussian Process Classification* (GPC). Penelitian berfokus membandingkan model berbasis GPC dan tiga jenis klasifikasi (*Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), dan *Naive Bayes*). Hasil yang didapatkan dari penelitian bahwa klasifikasi dengan model berbasis GP menghasilkan akurasi 81,97%.

Pada penelitian [8] membandingkan enam metode pembelajaran mesin, seperti *Logistic Regression*, *K-Nearest Neighbor*, *Support Vector Machine*, *Naïve Bayes*, *Decision Tree*, dan *Random Forest* dan menggunakan dua dataset, yaitu kuesioner *dataset* yang terdiri 18 pertanyaan dan PIMA *dataset*. Dihasilkan metode random forest memiliki akurasi yang lebih tinggi untuk kedua *dataset* dengan akurasi 94.1 % menggunakan kuesioner *dataset*.

Pada penelitian [9] merancang sistem klasifikasi dengan metode *Decision Tree* (J48) dan *Naïve Bayes* menggunakan *Pima Indians Diabetes Database* dari *National Institute of Diabetes*. Sistem yang dirancang melalui proses *preprocessing* dengan mengganti nilai yang hilang dan normalisasi nilai serta menggunakan teknik persentase *split* 70:30. Hasil penelitian dengan metode *Naïve Bayes* memperoleh akurasi 79.56% sedangkan *Decision Tree (J48)* memperoleh akurasi 76.95%.

Pada penelitian [10] menggunakan *repository* dari UCI *machine learning* dengan membandingkan metode *Decision Tree* (J48), *K-Nearest Neighbor*, *Random Forest*, *Support Vector Machine*. Dihasilkan metode Decision Tree (J48) memiliki akurasi 73.82% lebih tinggi dari ketiga metode lainya, namun hasil tersebut saat *dataset* belum dilakukan *preprocessing*. Setelah melakukan *preprocessing* pada *noise data*, metode *K-Nearest Neighbor* (k = 1) dan *Random Forest* memiliki akurasi lebih tinggi daripada metode lain dengan akurasi mencapai 100%.

Pada penelitian [11] membandingkan tiga metode, yaitu *Binary Logistic Regression, Multilayer Perceptron dan K-Nearest Neighbor*. Dataset yang dikumpulkan dari *multi-dimensional healthcare dataset* yang berisi seratus observasi dan tujuh atribut. Dari hasil penelitian didapatkan bahwa metode *K-Nearest Neighbor* memiliki akurasi 80%, *Binary Logistic Regression* memiliki akurasi 69%, dan *Multilayer Perceptron* memiliki akurasi 71%.

Penelitian [12] melakukan analisis terhadap beberapa teknik *preprocessing* dan mengidentifikasi teknik *preprocessing* yang memiliki performa lebih baik. Penelitian ini merancang model klasifikasi yang dapat mengkategorikan seseorang menderita diabetes atau tidak dengan menggunakan beberapa metode, seperti ID3, CART, C4.5 dan ANN. Hasil dari penelitian menunjukkan bahwa metode ANN memiliki akurasi lebih baik daripada ketiga metode Decision Tree dengan akurasi 81%.