Perbandingan Teknik Penanganan Missing Value Dalam Klasifikasi Penyakit Diabetes Menggunakan Metode Naïve Bayes

Alkea Harry Mauladha[1], Adiwijaya[2], Widi Astuti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]alkeaharrym@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]widiwdu@telkomuniversity.ac.id

Abstrak

Diabetes Melitus (DM) merupakan salah satu penyakit berbahaya di kalangan dewasa maupun remaja. Penyebab utama diabetes adalah peningkatan kadar gula dalam darah serta pola hidup tidak sehat juga merupakan salah satu pemicu seseorang terkena diabetes. Hal ini terjadi karena adanya gangguan kerja insulin yang tidak dapat mengubah kadar gula menjadi energi. Menurut *World Health Organization,* ada 400 juta lebih penderita diabetes di seluruh dunia. Dengan banyaknya angka penderita diabetes, dibutuhkan solusi untuk dapat mendeteksi penyakit diabetes secara cepat. Penelitian ini bertujuan untuk mendeteksi penderita diabetes menggunakan algoritma Naïve Bayes dengan membandingkan teknik penanganan *missing value,* yaitu menghapus *record* yang berisi atribut *missing value* dan *unsupervised imputation*. Naïve Bayes yang merupakan salah satu metode yang dianggap sebagai algoritma yang efektif yang digunakan untuk proses klasifikasi. Penelitian ini menggunakan *Dataset* Gula Karya Medika. Hasil penelitian menunjukkan dengan menggunakan teknik mengisi *missing value* dengan nilai *mean* dan *median* serta menggunakan 3-fold cross validation mendapatkan hasil terbaik dengan rata-rata *accuracy sebesar 74.9%, precision 90.64%, recall 66.57%, f-score 76.33%.*

Kata Kunci: diabetes, naïve bayes, klasifikasi, missing value, unsupervised imputation.

Abstract

Diabetes Mellitus (DM) is one of the dangerous diseases among adults and adolescents. The main cause of diabetes is increased levels of sugar in the blood as well as an unhealthy lifestyle is also one of the triggers of a person with diabetes. This occurs due to the disruption of insulin work that can not convert sugar levels into energy. According to the World Health Organization, there are over 400 million diabetics worldwide. With the large number of diabetics, a solution is needed to be able to detect diabetes disease quickly. This study aims to detect diabetics using naïve bayes algorithm by comparing missing value handling techniques, namely deleting records containing missing value and unsupervised imputation attributes. Naïve Bayes is one of the methods considered an effective algorithm used for the classification process. This study uses Medika's Sugar Dataset. The results showed by using the technique of filling missing values with mean and median values and using 3-fold cross validation get the best results with average accuracy of 74.9%, precision 90.64%, recall 66.57%, f-score 76.33%.

Keywords: diabetes, naïve bayes, classification, missing value, unsupervised imputation.

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%.

2.2 Diabetes

Diabetes merupakan salah satu penyakit tidak menular yang berbahaya dimana pankreas tidak dapat menghasilkan hormon insulin yang berguna untuk mengatur metabolisme karbohidrat, sehingga glukosa yang dikonsumsi tidak dapat mengalir ke sel tubuh untuk menghasilkan energi. *World Health Organization* [1] menyatakan diabetes adalah penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah, yang dari waktu ke waktu menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal, dan saraf. Gejala umum yang seseorang menderita diabetes, yaitu Poliuria (urine berlebih), Polyphagia (rasa lapar berlebih), Polidipsia (haus berlebih), berat badan naik atau turun tidak normal, penyembuhan luka tidak cepat, penglihatan kabur, dan kelelahan [9].

Penyakit diabetes terdiri dari tiga jenis, yaitu diabetes tipe 1, diabetes tipe 2, dan diabetes *gestasional*. Diabetes tipe 1 biasanya terjadi pada remaja atau anak-anak yang disebabkan oleh reaksi autoimun dimana sel-sel tidak dapat memproduksi insulin. Produksi urin berlebih (Poliuria), rasa lapar terus menerus, penurunan berat badan, dan kelelahan merupakan gejala diabetes tipe 1. Diabetes tipe 2 merupakan diabetes yang paling umum terjadi oleh penderita diabetes, tipe ini terjadi karena insulin tidak bekerja dengan baik (produksi insulin sedikit) di dalam tubuh. Diabetes tipe 2 dapat disebabkan karena kurangnya olahraga dan kelebihan berat badan, biasanya tipe ini terjadi pada orang dewasa dan lansia. Diabetes *gestasional* merupakan diabetes dengan tingkat glukosa darah tinggi (hiperglikemia) yang terjadi pada saat kehamilan. Penderita diabetes gestasional memungkinkan melahirkan anak dengan kelebihan berat badan dan beresiko menderita diabetes.

2.3 Penanganan *Missing Value*

Missing value merupakan suatu informasi yang tidak tersedia dalam sebuah kasus atau dataset. *Missing value* dapat terjadi karena informasi tidak ada atau tidak diberikan. Selain itu, *missing value* dapat terjadi dari non-respons parsial, penolakan menjawab pertanyaan, respons yang tidak dipahami, kehilangan data, melewatkan pertanyaan, dan alasan terkait [13]. Secara umum. terdapat tiga kategori data yang hilang [5], yaitu:

- *Missing Completely at Random* (MCAR) atau hilang sepenuhnya secara acak: Data hilang secara independen dari keduanya data yang diamati dan data yang tidak diamati. Misalnya, dalam survei siswa, jika mendapat tanggapan 5% hilang secara acak.
- *Missing at Rando*m (MAR) atau hilang secara acak: Data hilang secara independen dari data yang tidak teramati. Misalnya, jika kita mendapatkan 10% tanggapan survei yang hilang untuk siswa laki-laki dan 5% hilang untuk survei siswa perempuan.
- *Not Missing at Random* (NMAR) atau hilang secara tidak acak: Pengamatan yang hilang terkait dengan nilai data yang tidak diamati itu sendiri. Misalnya, jika menurunkan IPK mahasiswa, semakin tinggi tingkat respon survei.

Di dalam analisis data, *missing value* merupakan masalah yang wajar, namun *missing value* merupakan salah satu faktor yang dapat menurunkan performa [5]. Oleh karena itu, penanganan *missing value* menjadi salah satu teknik untuk meningkatkan performa dan juga mengatasi kekosongan data untuk melakukan model klasifikasi. Secara umum, ada beberapa metode umum yang dapat dilakukan untuk mengatasi *missing value*, seperti menghilangkan *record* yang berisi *missing value*, *unsupervised imputation*, *supervised imputation*.

Unsupervised imputation adalah teknik penanganan *missing value* tanpa pengawasan yang berarti tidak menggunakan atribut target sebagai parameter. *Unsupervised imputation* dapat menangani data dengan menggunakan metode sederhana, seperti pendekatan statistik dan *hot-deck imputation*. Dalam pendekatan statistik, mengisi data yang hilang dengan nilai *mean*, *median*, dan mode menjadi teknik yang paling umum digunakan. Hot-Deck imputation merupakan teknik yang mengisi nilai dengan contoh yang paling mirip dengan *missing value*. Ada beberapa contoh untuk mencari contoh yang paling mirip dengan *missing value*, seperti fungsi jarak yang dapat mengukur kesamaan atribut. Penggunaan fungsi jarak untuk atribut kontinu harus menggunakan Euclidean atau Manhattan. Sedangkan s*upervised imputation* adalah teknik penanganan *missing value* yang menggunakan atribut kelas sebagai parameter, umumnya digunakan dengan supervised algorithms. Teknik pengisian *missing value* dilakukan dengan melakukan metode klasifikasi seperti CLIP$, Naïve Bayes, C4.5 [14].

Selain itu, ada beberapa teknik penanganan *missing value* yang dikemukakan [13], seperti respon pembobotan, analisis nilai, mengganti dengan nilai *default*, *multiple imputation*, *single imputation*, *maximum likelihood estimation*. Sedangkan menurut [5] ada beberapa teknik penanganan *missing value*, seperti s*ingle imputation, multiple imputation, predictive mean matching, logistic regression, polytomous logistic regression, linear discriminant analysis, classification and regression tree, Bayesian linear regression,* dan *amelia*.

2.4 Naïve Bayes

Naïve Bayes adalah salah satu algoritma yang digunakan untuk klasifikasi probabilistik di dalam data mining. Naïve Bayes dikenalkan oleh ilmuwan asal Inggris bernama Thomas Bayes. Algoritma Naive Bayes adalah pengklasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menghitung frekuensi dan kombinasi nilai dalam kumpulan data tertentu [15]. Naïve Bayes mendefinisikan seluruh atribut independen atau tidak saling berkaitan satu sama lain. Metode ini dianggap sebagai metode yang bagus dan efisien untuk klasifikasi [4]. Definisi lain mengatakan algoritma Naïve Bayes menggunakan teorema Bayes dan mengasumsikan seluruh atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada atribut kelas [16].

Teorema Naïve Bayes sering digunakan untuk melakukan klasifikasi sederhana maupun kompleks [17]. Keuntungan menggunakan metode Naïve Bayes adalah model ini mudah dibuat tanpa estimasi parameter berulang yang rumit yang membuatnya sangat berfungsi untuk kumpulan data yang sangat besar. Naive Bayes kerap berhasil dengan baik dan banyak digunakan karena sering mengungguli metode klasifikasi yang lebih canggih [18].

Dalam teorema Bayes, untuk menghitung probabilitas posterior dinyatakan sebagai [8]:

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)} \quad (1)$$

Dimana: P (c | x) adalah probabilitas posterior.
P (c) adalah prior probabilitas kelas.
P (x | c) adalah *likelihood*.
P (x) adalah prior probabilitas prediktor.

Untuk menghitung probabilitas data kontinu atau numerik dapat menggunakan persamaan berikut [19]:

$$P(x \mid c) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Dimana: μ adalah *mean* yang menyatakan rata-rata semua atribut.
σ adalah standar deviasi.
x adalah atribut (*variable*).
c adalah kelas (*class*).
π adalah nilai konstanta pi 3.1416.
e adalah bilangan euler, nilai konstanta 2.7183.

Untuk menghitung nilai rata-rata dan standar deviasi untuk atribut numerik menggunakan rumus sebagai berikut [20]:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (3)$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{0.5} \quad (4)$$

2.5 Evaluasi Performa

Evaluasi performa terhadap sistem klasifikasi adalah hal yang penting dilakukan. Hasil performa dari suatu sistem menunjukkan sejauh mana sistem dapat mengklasifikasi data. Untuk mengevaluasi performa pada sistem yang dibuat, sistem menggunakan empat jenis pengujian. yaitu *Accuracy*, *Recall*, *Precision*, dan *F-Score* dengan metode *Confusion Matrix*.

A. Confusion Matrix

Confusion Matrix adalah sebuah metode yang digunakan untuk menghitung performa akurasi pada data mining.

Tabel 1. *Confusion Matrix*

		Kelas Prediksi	
Kelas Aktual		Positif	Negatif
	Positif	***True Positive* (TP)**	***False Negative* (FN)**
	Negatif	***False Positive* (FP)**	***True Negative* (TN)**

Berdasarkan Tabel 1, ada empat istilah sebagai karakterisasi *Confusion Matrix*, yaitu:

- *True Positive* (TP) menunjukkan data diabetes positif (Kelas 1) yang diklasifikasikan dengan benar.
- *True Negative* (TN) menunjukkan data diabetes negatif (Kelas 0) yang diklasifikasikan dengan benar.
- *False Negative* (FN) menunjukkan data diabetes positif (Kelas 1) yang diklasifikasikan sebagai negatif.
- *False Positive* (FP) menunjukkan data diabetes negatif (Kelas 0) yang diklasifikasikan sebagai positif.

B. Accuracy

Accuracy menunjukkan rasio prediksi data yang diklasifikasikan secara benar dengan seluruh data. Accuracy dapat dihitung dengan persamaan berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

C. Recall

Recall merupakan rasio prediksi *true positive* dibandingkan dengan keseluruhan data aktual positif. *Recall* dapat dihitung dengan persamaan berikut:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

D. Precision

Precision merupakan rasio prediksi data *true positive* dengan seluruh data yang diprediksi positif. *Precision* dapat dihitung dengan persamaan berikut:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

E. F– Score

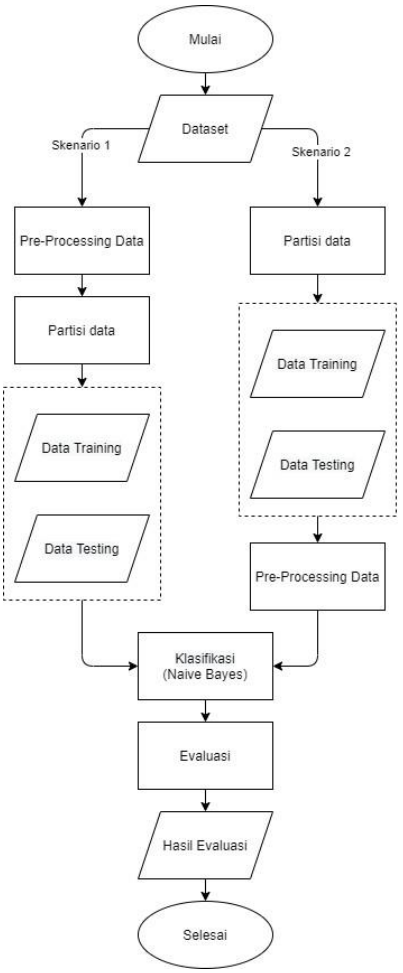
F-Score merupakan penggabungan antara cara *recall* dan *precision* dimana *F-Score* dapat dikatakan sebagai *harmonic mean* antara *recall* dan *precision*. *F-Score* dapat dihitung dengan persamaan berikut:

$$F - \text{Score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

3. Sistem yang Dibangun

3.1 Rancangan Sistem

Berikut adalah rancangan sistem yang akan dibuat dalam melakukan klasifikasi penyakit diabetes.



Gambar 1. *Flowchart* Rancangan Sistem

3.2 Dataset

Pada penelitian ini, *dataset* yang digunakan adalah *Dataset* Gula Karya Medika. *Dataset* Gula Karya Medika juga digunakan pada penlitian [21]. *Dataset* memiliki lima atribut dan satu atribut kelas juga memiliki 470 *record* yang terdiri dari 278 pria dan 192 wanita. *Dataset* menyatakan bahwa 290 *record* menderita diabetes dan 180 *record* tidak menderita diabetes. Contoh *dataset* yang akan digunakan dapat dilihat pada Tabel 2.

Tabel 2. *Dataset* Gula Karya Medika

Dataset	Jumlah *Record*	Jumlah Atribut	Tipe Data	Distribusi Atribut Kelas
Dataset Gula Karya Medika	470	6	Numerik	Kelas 0 - Non-Diabetes (180) Kelas 1 – Diabetes (290)

Tabel 3 merupakan contoh Dataset Gula Karya Medika.

Tabel 3. Contoh *Dataset*

No.	*Glucose*	*Gender*	*Blood Pressure*	BMI	Usia	Class

1	157	1	80	21.6	49	1
2	158	1	88	32.9	50	1
…	…	…	…	…	…	…
469	82	0	70	29.6	43	0
470	75	1	100	21.7	45	0

3.3 *Preprocessing*

Preprocessing data merupakan tahap atau proses untuk memanipulasi data sebelum melakukan proses klasifikasi. *preprocessing* adalah salah satu proses untuk meningkatkan kualitas dataset yang akan digunakan dan meningkatkan performa klasifikasi. Pada *Dataset* Gula Karya Medika ditemukan *missing value* untuk beberapa atribut *Blood Pressure*, BMI, dan Usia yang digambarkan dengan nilai "0" (bukan angka) atau "*NaN*". *Missing value* dapat mengakibatkan kualitas data dan performa klasifikasi berkurang. Teknik penanganan *missing value* terdiri dari menghapus *record* yang berisi atribut *missing value* dan *unsupervised imputation*. *Unsupervised imputation* berfungsi untuk menangani data yang hilang dengan metode yang sangat sederhana seperti imputasi rata-rata hingga metode statistik berdasarkan estimasi parameter [14]. Waktu penanganan *preprocessing* dibagi menjadi dua skenario, yaitu sebelum partisi data dan sesudah partisi data. Menurut penelitian [22], dijelaskan bahwa teknik *unsupervised imputation* lebih baik digunakan sebelum partisi data karena tidak menimbulkan biaya komputasi yang tinggi. Sedangkan menurut penelitian [23], saat melakukan *preprocessing* data sebaiknya dilakukan setelah partisi data untuk mencegah kebocoran pada data. Teknik *preprocessing* yang digunakan pada data menggunakan teknik penanganan *missing value* dan normalisasi data.

Teknik penanganan *missing value* meliputi:

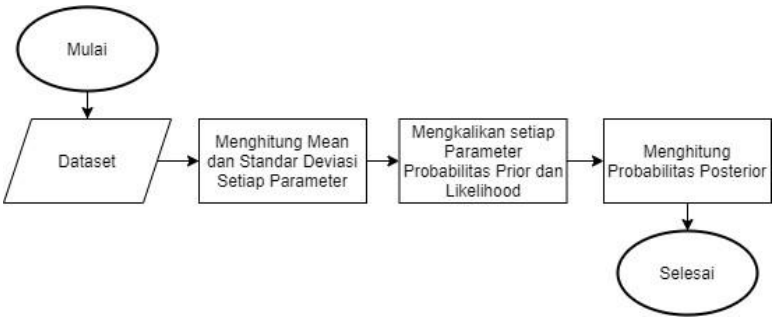
- Menghapus *record* yang berisi atribut *missing value.*
- Mengisi missing value dengan nilai 0.
- Mengisi *missing value* dengan rata-rata atribut yang sesuai (*mean)*.
- Mengisi *missing value* dengan nilai tengah atribut yang sesuai (*median).*
- Menggunakan *K-Nearest Neighbor* untuk mengisi *missing value* (*KNNImputer*).

Dengan adanya kelima strategi yang digunakan, penelitian ini dapat mengetahui *preprocessing* yang layak digunakan untuk mengatasi *missing value.*

3.4 Partisi Data

Partisi data merupakan tahap untuk memisahkan *Dataset* Gula Karya Medika menjadi dua bagian, yaitu *data training* dan *data testing* dengan menggunakan teknik *k-fold cross validation*. *Data training* merupakan data yang berguna untuk proses pelatihan oleh model algoritma dan *data testing* digunakan sebagai data validasi untuk mengetahui performa dari data yang sudah dilatih oleh model algoritma. *Cross validation* adalah pendekatan alternatif partisi tetap. Dalam *cross validation*, beberapa partisi dihasilkan (*data training* dan *data testing*), memungkinkan setiap sampel data digunakan beberapa kali. Metode *k-fold cross validation* melibatkan pemisahan secara acak kumpulan sampel menjadi serangkaian lipatan yang berukuran sama, di mana k menunjukkan jumlah lipatan [24]. Namun, pada pengujian ini menggunakan partisi data *Stratified K-Fold* dimana setiap *fold* didistribusikan secara merata, sehingga mengandung jumlah label atau kelas target yang kira-kira sama dengan kumpulan data asli [24]. Dalam pengujian ini k yang digunakan 3 dan 5. Penggunaan k = 3 dan k = 5 bertujuan untuk membandingkan jumlah sampel *data training* dan *data testing* yang berbeda. Menurut [25], banyak kasus dengan model BN cukup menggunakan k = 5 untuk menghemat waktu dan masalah kompleksitas komputasi. Di sisi lain, k = 3 (314 *record data training* dan 156 *record data testing*) juga dipilih karena memiliki jumlah data testing yang lebih besar daripada k = 5.

3.5 Klasifikasi



Gambar 2. *Flowchart* Klasifikasi *Gaussian Naïve Bayes*

Klasifikasi merupakan tahap untuk memprediksi kelas target secara akurat untuk setiap kasus dalam *dataset* yang sudah melalui tahap *preprocessing data* dan partisi data. Metode klasifikasi yang digunakan pada sistem yang dibuat adalah Gaussian Naïve Bayes. Penelitian ini menggunakan metode klasifikasi *Gaussian Naïve Bayes* karena atribut pada dataset yang digunakan berisi tipe data numerik. Keluaran dari proses klasifikasi akan dites untuk mengetahui nilai performa.

3.6 Evaluasi Performa

Setelah melakukan lima strategi *preprocessing*, yaitu menghapus *record* menghapus *record* yang berisi atribut *missing value*, mengisi *missing value* dengan nilai 0, mengisi *missing value* dengan *mean*, mengisi *missing value* dengan *median*, dan menggunakan *K-Nearest Neighbor* untuk mengisi *missing value*. Selanjutnya melakukan klasifikasi Naïve Bayes dengan strategi *preprocessing* yang berbeda. Evaluasi performa digunakan untuk mengevaluasi sistem yang telah dibuat menggunakan *Confusion Matrix*. Proses ini akan menghitung performa *accuracy, precision, recall, dan f-score*.

3.7 Hasil Evaluasi

Hasil evaluasi merupakan nilai dari *accuracy*, *precision*, *recall, dan f-score* yang telah dihitung pada tahap evaluasi performa. Hasil evaluasi menunjukkan seberapa akurat sistem klasifikasi yang telah dibuat, sehingga dapat menjadi acuan untuk dapat mendeteksi penderita diabetes.

4. Evaluasi

4.1 Hasil Pengujian

Bagian ini akan membahas hasil dari metode yang diusulkan. Pengujian ini menggunakan Dataset Gula Karya Medika dengan menerapkan teknik penanganan *missing value* dilanjutkan dengan partisi data menggunakan *Stratified K-Fold Cross Validation* kemudian melakukan klasifikasi dengan *Gaussian Naïve Bayes*.

Pada pengujian yang diusulkan terdapat beberapa skenario pengujian yang akan dilakukan dengan membandingkan lima macam penanganan *missing value*, yaitu menghapus *record* yang berisi atribut *missing value,* mengisi *missing value* dengan *mean* (nilai rata-rata yang ada di atribut), mengisi *missing value* dengan nilai 0, *median*, menggunakan *k-nearest neighbor* untuk mengisi *missing value.* Pertama, pengujian dilakukan dengan melakukan teknik *preprocessing*, kemudian dilanjutkan dengan partisi data menggunakan *k-fold cross validation*. Kedua, pengujian dilakukan dengan partisi data, kemudian dilanjutkan dengan teknik *preprocessing*. Untuk *k-fold cross validation* menggunakan k = 3 dan k = 5. Hasil skenario pengujian akan dilakukan analisis dan dibandingkan yang bertujuan untuk mengetahui performa terbaik. Skenario pengujian dapat dilihat pada Tabel 4.

Tabel 4. Skenario Pengujian

No.	Waktu Penanganan *Missing Value*	Teknik Penanganan *Missing Value*	K-Fold
1	Sebelum Partisi Data	*Dropping record*	3
2			5
3		Diisi dengan Nilai 0	3
4			5
5		Diisi dengan Nilai *Mean*	3
6			5

7		Diisi dengan Nilai *Median*	3
8			5
9		Diisi dengan *KNNImputer*	3
10			5
11	Setelah Partisi Data	*Dropping record*	3
12			5
13		Diisi dengan Nilai 0	3
14			5
15		Diisi dengan Nilai *Mean*	3
16			5
17		Diisi dengan Nilai *Median*	3
18			5
19		Diisi dengan *KNNImputer*	3
20			5

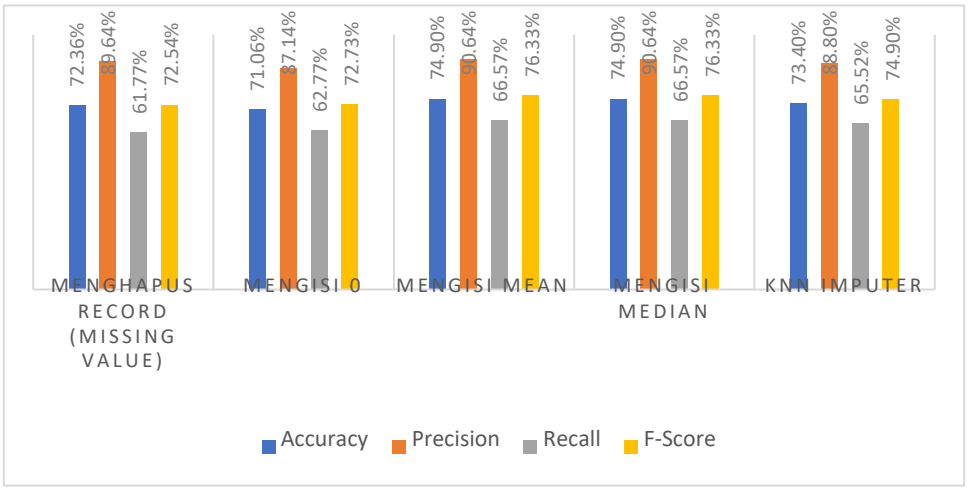
4.1.1 Pengujian Teknik *Preprocessing* sebelum Partisi Data

Pada pengujian pertama dilakukan pengujian dengan *Gaussian Naïve Bayes* dengan membandingkan teknik penanganan *missing value*. Dalam pengujian ini teknik *preprocessing* dilakukan setelah partisi data berupa *k-fold cross validation*, k = 3 dan k = 5. Dimana *k-fold cross validation* merupakan metode yang membagi dua data yaitu, *data training* dan *data testing*. Dalam *k-fold cross validation*, *data training* dan *data testing* akan dibagi menjadi *k* kali lipat dimana dataset dibagi menjadi *k* - 1 subset *data training* dan 1 subset *data testing*. Hasil pada pengujian pertama dengan k = 3 didapat bahwa mengisi dengan nilai *mean* dan *median* memperoleh hasil performa terbaik dengan rata-rata akurasi sebesar 74.9% Hasil pengujian dapat dilihat pada Tabel 4.

Tabel 5. Performa Teknik *Preprocessing* sebelum 3-*fold*

Preprocessing	***Accuracy***	***Precision***	***Recall***	***F-Score***
Menghapus *Record* (*Missing Value*)	**72.36%**	**89.64%**	**61.77%**	**72.54%**
Mengisi 0	**71.06%**	**87.14%**	**62.77%**	**72.73%**
Mengisi *Mean*	**74.9%**	**90.64%**	**66.57%**	**76.33%**
Mengisi *Median*	**74.9%**	**90.64%**	**66.57%**	**76.33%**
KNN Imputer	**73.4%**	**88.8%**	**65.52%**	**74.9%**

Visualisasi hasil performa dapat dilihat pada Gambar 3.



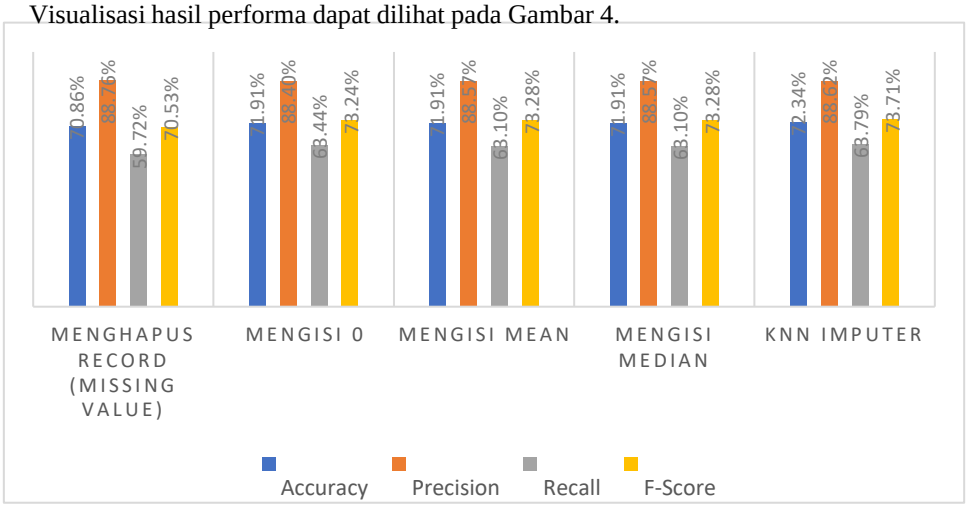
Gambar 3. Performa Teknik *Preprocessing* sebelum 3-*fold*

Hasil pada pengujian pertama dengan k = 5 didapat bahwa mengisi nilai dengan teknik *KNNImputer* memperoleh hasil performa terbaik dengan rata-rata akurasi sebesar 72.34% Hasil rincian pengujian dapat dilihat pada Tabel 5.

Tabel 6. Performa Teknik *Preprocessing* sebelum 5-*fold*

Preprocessing	***Accuracy***	***Precision***	***Recall***	***F-Score***
Menghapus *Record* (*Missing Value*)	**70.86%**	**88.76%**	**59.72%**	**70.53%**
Mengisi 0	**71.91%**	**88.4%**	**63.44%**	**73.24%**
Mengisi *Mean*	**71.91%**	**88.57%**	**63.1%**	**73.28%**
Mengisi *Median*	**71.91%**	**88.57%**	**63.1%**	**73.28%**
KNN Imputer	**72.34%**	**88.62%**	**63.79%**	**73.71%**

Visualisasi hasil performa dapat dilihat pada Gambar 4.



Gambar 4. Performa Teknik *Preprocessing* sebelum 5-*fold*

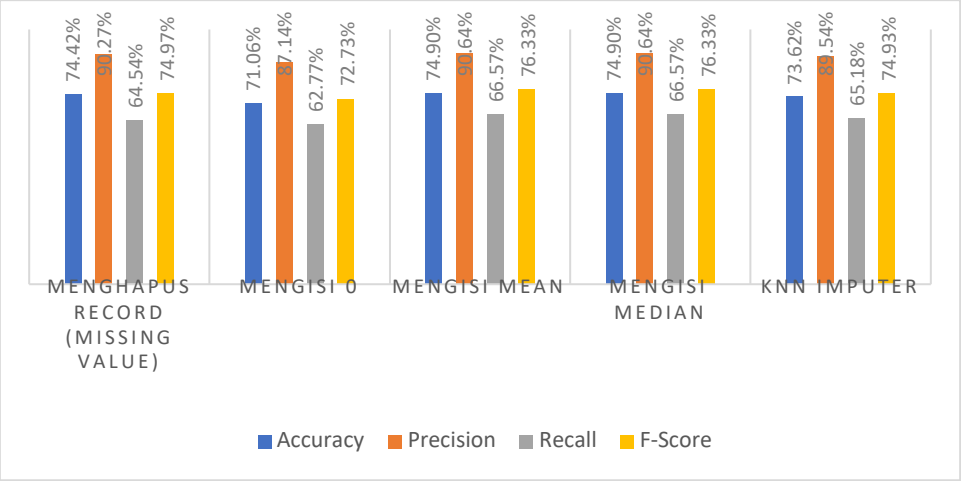
4.1.2 Pengujian Teknik *Preprocessing* setelah Partisi Data

Pada pengujian kedua dilakukan dengan metode dan teknik penanganan *missing value* yang sama dengan pengujian pertama. Namun, pada pengujian kedua ini teknik *preprocessing* dilakukan setelah partisi data menggunakan *k-fold cross validation*, k = 3 dan k = 5. Hasil pada pengujian kedua dengan k = 3 didapat bahwa mengisi nilai *missing value* dengan *median* dan *mean* memperoleh hasil performa terbaik dengan rata-rata akurasi sebesar 74.9%. Hasil rincian pengujian dapat dilihat pada Tabel 6.

Tabel 7. Performa Teknik *Preprocessing* setelah 3-*fold*

Preprocessing	***Accuracy***	***Precision***	***Recall***	***F-Score***
Menghapus *Record* (*Missing Value*)	**74.42%**	**90.27%**	**64.54%**	**74.97%**
Mengisi 0	**71.06%**	**87.14%**	**62.77%**	**72.73%**
Mengisi *Mean*	**74.9%**	**90.64%**	**66.57%**	**76.33%**
Mengisi *Median*	**74.9%**	**90.64%**	**66.57%**	**76.33%**
KNN Imputer	**73.62%**	**89.54%**	**65.18%**	**74.93%**

Visualisasi hasil performa dapat dilihat pada Gambar 5.



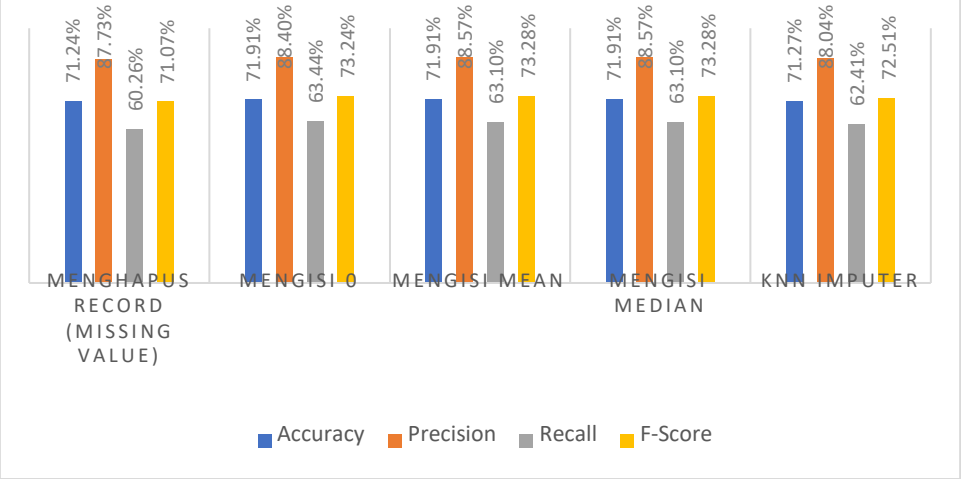
Gambar 5. Performa Teknik *Preprocessing* setelah 3-*fold*

Hasil pada pengujian kedua dengan k = 5 didapat bahwa mengisi nilai *missing value* dengan nilai 0, *median*, dan *mean* memperoleh hasil performa terbaik dengan akurasi sebesar rata-rata 71.91% Hasil rincian pengujian dapat dilihat pada Tabel 7.

Tabel 8. Performa Teknik *Preprocessing* setelah 5-*fold*

Preprocessing	***Accuracy***	***Precision***	***Recall***	***F-Score***
Menghapus *Record* (*Missing Value*)	**71.24%**	**87.73%**	**60.26%**	**71.07%**
Mengisi 0	**71.91%**	**88.4%**	**63.44%**	**73.24%**
Mengisi *Mean*	**71.91%**	**88.57%**	**63.1%**	**73.28%**
Mengisi *Median*	**71.91%**	**88.57%**	**63.1%**	**73.28%**
KNN Imputer	**71.27%**	**88.04%**	**62.41%**	**72.51%**

Visualisasi hasil performa dapat dilihat pada Gambar 6.



Gambar 6. Performa Teknik *Preprocessing* setelah 5-*fold*

4.2 Analisis Hasil Pengujian

Pada penelitian ini dilakukan perhitungan evaluasi performa terhadap setiap teknik penanganan *missing value* dan partisi data. Penelitian ini menggunakan *accuracy*, *precision*, *recall*, dan *f-score* untuk membandingkan performa model penelitian.

Pada pengujian pertama dimana lima teknik *preprocessing* dilakukan sebelum partisi data *k-fold cross validatio*n dengan menggunakan k = 3 dan k = 5. Hasil pengujian 3-*fold cross validation* dengan teknik menghapus *record* yang terdapat *missing value* mendapatkan performa rata-rata *accuracy* sebesar 72,36%, *precision* 89.64%, *recall* 61.77%, *f-score* 72.54%. Dengan mengisi nilai 0 mendapatkan hasil performa rata-rata *accuracy* sebesar 71.06%, *precision* 87.14 %, *recall* 62.77 %, *f-score* 72.73 %. Dengan teknik *KNNImputer* diperoleh performa rata-rata *accuracy* sebesar 73.4%, *precision* 88.8%, *recall* 65.52%, *f-score* 74.9%. Sedangkan dengan mengisi nilai *mean* dan *median* dari setiap atribut memperoleh hasil performa terbaik dengan rata-rata *accuracy* sebesar 74.9%, *precision* 90.64%, *recall* 66.57%, *f-score* 76.33%.

Hasil pengujian 5-*fold cross validation* dengan teknik menghapus *record* yang terdapat *missing value* mendapatkan performa rata-rata *accuracy* sebesar 70.86%, *precision* 88.76%, *recall* 59.72%, *f-score* 70.53%. Selanjutnya, dengan mengisi missing value dengan nilai *mean* dan *median* mendapatkan hasil rata-rata *accuracy* sebesar 71.91%, *precision* 88.57%, *recall* 63.1%, *f-score* 73.28%. Dengan mengisi nilai 0 rata-rata *accuracy* sebesar 71.91%, *precision* 88.4%, *recall* 63.44%, *f-score* 73.24%. Sedangkan dengan teknik *KNNImputer* mendapatkan performa terbaik dengan rata-rata *accuracy* sebesar 72.34%, *precision* 88.62%, *recall* 63.79%, *f-score* 73.71%. Berdasarkan hasil pengujian pertama menggunakan teknik *preprocessing* sebelum partisi data nilai k = 3 memiliki rata-rata performa yang lebih baik daripada k = 5 pada Dataset Gula Karya Medika, selain itu penanganan *missing value* dengan mengisi nilai *mean* dan *median* layak digunakan pada Dataset Gula Karya Medika.

Pada pengujian kedua dimana lima teknik *preprocessing* dilakukan setelah partisi data dengan *3-fold cross validation*. Hasil pengujian dengan 3-*fold cross validation* menggunakan teknik menghapus *record* yang terdapat *missing value* mendapat hasil performa rata-rata *accuracy* sebesar 74.42%, *precision* 90.27%, *recall* 64.54%, *f-score* 74.97%. Dengan mengisi nilai 0 rata-rata *accuracy* sebesar 71.06%, *precision* 87.14 %, *recall* 62.77 %, *f-score* 72.73 %. Selanjutnya, dengan menggunakan teknik *KNNImputer* diperoleh performa rata-rata *accuracy* sebesar 73.62%, *precision* 89.54%, *recall* 65.18%, *f-score* 74.93%. Sedangkan mengisi nilai dengan nilai *median* dan *mean* memperoleh hasil terbaik untuk 3-*fold cross validation*, dengan rata-rata *accuracy* sebesar 74.9%, *precision* 90.64%, *recall* 66.57%, *f-score* 76.33%.

Hasil pengujian menggunakan partisi data 5-*fold cross validation* dengan menghapus *record* mendapat rata-rata *accuracy* sebesar 71.24%, *precision* 87.73%, *recall* 60.26%, *f-score* 71.07%. Dengan mengisi nilai 0 rata-rata *accuracy* sebesar 71.91%, *precision* 88.4%, *recall* 63.44%, *f-score* 73.24%. Kemudian dengan teknik *KNNImputer* mendapatkan rata-rata *accuracy* sebesar 71.27%, *precision* 88.04%, *recall* 62.41%, *f-score* 72.51%. Sedangkan mengisi *missing value* dengan nilai *mean* dan *median* mendapat hasil *accuracy* sebesar 71.91%, *precision* 88.57%, *recall* 63.1%, *f-score* 73.28%. Berdasarkan hasil pengujian kedua, nilai k = 3 memiliki rata-rata performa yang lebih baik daripada k = 5 pada Dataset Gula Karya Medika dan teknik penanganan *missing value* dengan mengisi nilai *mean* dan *median* merupakan pilihan yang baik digunakan pada Dataset Gula Karya Medika.

Dengan menggunakan algoritma *gaussian naïve bayes* untuk mengklasifikasi penderita diabetes dan membandingkan skenario pengujian yang sudah dilakukan menunjukkan bahwa teknik penanganan *missing value* dengan mengisi nilai *mean* dan *median* mampu meningkatkan performa yang signifikan pada klasifikasi. Juga penggunaan *k-fold cross validation* k = 3 memperoleh performa lebih baik daripada k = 5. Teknik penanganan *missing value* dengan mengisi nilai *mean* dan *median* menghasilkan performa yang sama dipengaruhi oleh partisi data yang dilakukan.

Selain itu, jika membandingkan teknik penanganan *missing value* sebelum maupun setelah partisi data dapat dikatakan bahwa teknik *unsupervised imputation* (*mean, median, dan KNNImputer*) memiliki rata-rata akurasi lebih baik dengan mengisi nilai sebelum partisi data, sedangkan menghapus *record* yang berisi *missing value* memiliki rata-rata akurasi lebih baik dengan menghapus setelah partisi data.

5. Kesimpulan

Berdasarkan hasil beberapa skenario pengujian yang dilakukan untuk menganalisis lima teknik penanganan *missing value* dan partisi data pada klasifikasi diabetes menggunakan algoritma *gaussian naïve bayes*, dapat disimpulkan bahwa metode gaussian naïve bayes mampu memberikan hasil yang berbeda tergantung dengan skenario pengujian yang digunakan. Performa terbaik dihasilkan dengan teknik mengisi *missing value* dengan nilai *mean* dan *median* dari masing-masing atribut dan menggunakan 3-*fold cross validation*. Performa terbaik menghasilkan rata-rata *accuracy* sebesar 74.9%, *precision* 90.64%, *recall* 66.57%, *f-score* 76.33%.

Selain itu, pada pengujian ini penggunaan teknik *unsupervised imputation* memiliki rata-rata performa lebih baik dengan mengisi nilai sebelum partisi data, sedangkan menghapus *record* yang berisi *missing value* memiliki rata-rata performa lebih baik dengan menghapus setelah partisi data. Pada Dataset Gula Karya Medika penggunaan *stratified k-fold cross validation* menggunakan k = 3 mendapat performa yang lebih baik dibandingkan k = 5.

Kendati demikian, penelitian ini masih dapat dikembangkan dan dilanjutkan. Oleh karena itu, saran untuk penelitian selanjutnya adalah dengan menggunakan teknik penanganan *missing value* dan teknik *preprocessing* yang lebih canggih, ataupun menggunakan metode klasifikasi yang berbeda. Diharapkan rancangan sistem yang dibangun dapat digunakan untuk memprediksi atau mendiagnosis penyakit lain.

Referensi

[1] "Diabetes." https://www.who.int/health-topics/diabetes (diakses Nov 22, 2020).

[2] "Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes (diakses Nov 22, 2020).

[3] "International Diabetes Federation - Facts & figures." https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html (diakses Nov 22, 2020).

[4] D. Sisodia dan D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, hal. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[5] S. I. Khan dan A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00313-w.

[6] Kumari V. Anuja dan Chitra R., "Classification Of Diabetes Disease Using Support Vector Machine," *Int. J. Eng. Res. Appl.*, vol. 3, no. 2, hal. 1797–1801, 2013.

[7] M. Maniruzzaman *et al.*, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, hal. 23–34, 2017, doi: 10.1016/j.cmpb.2017.09.004.

[8] N. P. Tigga dan S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Comput. Sci.*, vol. 167, no. 2019, hal. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[9] A. Iyer, J. S, dan R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, hal. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[10] J. P. Kandhasamy dan S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, no. C, hal. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182.

[11] S. Selvakumar, K. S. Kannan, dan S. Gothainachiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques," *Int. J. Stat. Syst.*, vol. 12, no. 2, hal. 183–188, 2017, [Daring]. Tersedia pada: http://www.ripublication.com.

[12] D. Venkata Vara Prasad, L. Venkataramana, P. Balasubramanian, B. Priyankha, S. Rajagopal, dan R.

Dattuluri, "An efficient pre-processing method for improved classification of diabetics using decision tree and artificial neural network," *AIP Conf. Proc.*, vol. 2161, no. October, 2019, doi: 10.1063/1.5127648.

[13] S. Gorard, "Handling missing data in numeric analyses," *Int. J. Soc. Res. Methodol.*, vol. 23, no. 6, hal. 651–660, 2020, doi: 10.1080/13645579.2020.1729974.

[14] A. Farhangfar, L. A. Kurgan, dan W. Pedrycz, "Experimental analysis of methods for imputation of missing values in databases," *Intell. Comput. Theory Appl. II*, vol. 5421, hal. 172, 2004, doi: 10.1117/12.542509.

[15] G. Dimitoglou, J. A. Adams, dan C. M. Jim, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," *J. Comput.*, vol. 4, no. 8, hal. 1–9, 2012.

[16] M. S. S. S. Tina R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *Int. J. Comput. Sci. Appl.*, vol. 6, no. 2, hal. 256–261, 2013, doi: 10.18201/ijisae.2019252786.

[17] H. Das, B. Naik, dan H. S. Behera, *Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach*, vol. 710, no. Dmd. Springer Singapore, 2018.

[18] "Naive Bayesian." http://www.saedsayad.com/naive_bayesian.htm (diakses Nov 22, 2020).

[19] H. Kamel, D. Abdulah, dan J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *Proc. 5th Int. Eng. Conf. IEC 2019*, hal. 165–170, 2019, doi: 10.1109/IEC47844.2019.8950650.

[20] A. Fadlil, I. Riadi, dan S. Aji, "DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 8, hal. 42–50, 2017, doi: 10.14569/ijacsa.2017.080806.

[21] G. A. B. Suryanegara, Adiwijaya, dan M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, hal. 114–122, 2021.

[22] B. C. Jaeger, N. J. Tierney, dan N. R. Simon, "When to Impute? Imputation before and during cross-validation," hal. 1–23, 2020, [Daring]. Tersedia pada: http://arxiv.org/abs/2010.00718.

[23] S. Kaufman, S. Rosset, C. Perlich, dan O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, hal. 556–563, 2012, doi: 10.1145/2382577.2382579.

[24] C. V. García-Mendoza, O. J. Gambino, M. G. Villarreal-Cervantes, dan H. Calvo, "Evolutionary optimization of ensemble learning to determine sentiment polarity in an unbalanced multiclass corpus," *Entropy*, vol. 22, no. 9, hal. 1–19, 2020, doi: 10.3390/e22091020.

[25] B. G. Marcot dan A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Comput. Stat.*, vol. 36, no. 3, hal. 2009–2031, 2021, doi: 10.1007/s00180-020-00999-9.