Jurnal Kajian Komputasi Inovatif Vol.17 No.2, Februari 2026

IMPLEMENTASI ALGORITMA *NAÏVE BAYES* DALAM KLASIFIKASI PENYAKIT DIABETES

Arini Diah Vita Loka[1], Wawan Joko Pranoto[2], Taghfirul Azhima Yoga Siswa[3]
Muhammadiyah Kalimantan Timur
E-mail: 2111102441169@umkt.ac.id[1], wjp337@umkt.ac.id[2], tay758@umkt.ac.id[3]

Abstrak

Penyakit Diabetes Mellitus merupakan salah satu ancaman kesehatan global yang memerlukan mekanisme deteksi dini yang presisi untuk meminimalisir risiko komplikasi kronis. Namun, efektivitas model klasifikasi seringkali terhambat oleh karakteristik data rekam medis yang tidak ideal. Penelitian ini bertujuan untuk menganalisis karakteristik data dan mengevaluasi kinerja algoritma Gaussian Naïve Bayes dalam mengklasifikasikan risiko diabetes pada pasien di Rumah Sakit Dirgahayu Samarinda. Metodologi yang diusulkan mencakup rangkaian pra-pemrosesan data yang komprehensif, dimulai dari penggunaan mean imputation untuk menjaga integritas data yang kosong, diikuti dengan transformasi Z-Score guna menormalisasi skala fitur klinis agar proses perhitungan probabilitas pada algoritma Naïve Bayes menjadi lebih objektif. Untuk mengatasi bias prediksi akibat ketimpangan data, diterapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) pada data latih. Selanjutnya, dilakukan eksperimen sistematis dengan membandingkan tiga skenario rasio pembagian data (split data), yaitu 90:10, 80:20, dan 70:30. Hasil penelitian mengonfirmasi bahwa integrasi algoritma Naïve Bayes dengan metode SMOTE mampu memberikan peningkatan performa yang stabil. Skenario pembagian data dengan rasio 80:20 teridentifikasi sebagai model yang paling optimal, dengan perolehan nilai Akurasi sebesar 83,19%, Presisi 79,17%, Recall 82,61%, dan F1-Score 80,85%. Kesimpulan dari penelitian ini menunjukkan bahwa pendekatan pra-pemrosesan yang tepat dan penyeimbangan data secara sintetis terbukti efektif dalam meningkatkan reliabilitas diagnosa komputasional, sehingga dapat diandalkan sebagai instrumen pendukung keputusan medis bagi tenaga profesional di rumah sakit.

Kata Kunci : Diabetes Mellitus, Gaussian Naïve Bayes, SMOTE, Imbalanced Data, Klasifikasi Medis.

Abstract

Diabetes Mellitus is one of the major global health threats that requires a precise early detection mechanism to minimize the risk of chronic complications. However, the effectiveness of classification models is often constrained by the non-ideal characteristics of medical record data. This study aims to analyze data characteristics and evaluate the performance of the Gaussian Naïve Bayes algorithm in classifying diabetes risk among patients at Dirgahayu Hospital, Samarinda. The proposed methodology involves a comprehensive data preprocessing pipeline, beginning with mean imputation to preserve the integrity of missing values, followed by Z-Score transformation to normalize the scale of clinical features and ensure objective probability calculations within the Naïve Bayes algorithm. To mitigate prediction bias caused by data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data. Furthermore, systematic experiments are conducted by comparing three data splitting scenarios, namely 90:10, 80:20, and 70:30. The experimental results confirm that integrating the Naïve Bayes algorithm with SMOTE leads to a stable improvement in

classification performance. Among the evaluated scenarios, the 80:20 data split is identified as the optimal model, achieving an accuracy of 83.19%, precision of 79.17%, recall of 82.61%, and an F1-score of 80.85%. The findings indicate that appropriate preprocessing strategies and synthetic data balancing are effective in enhancing the reliability of computational diagnosis, thereby supporting medical decision-making processes for healthcare professionals in hospital environments.
Keywords: *Diabetes Mellitus, Gaussian Naïve Bayes, SMOTE, Imbalanced Data, Classification.*

1. PENDAHULUAN

Penyakit Diabetes Mellitus (DM) merupakan krisis kesehatan global yang serius dan menjadi salah satu penyebab utama kematian. International Diabetes Federation (IDF) memproyeksikan jumlah penderita diabetes di dunia akan mencapai 783 juta jiwa pada tahun 2045, menunjukkan peningkatan signifikan setiap tahunnya (Arrayyan et al., 2024). Deteksi dini DM sangat penting untuk mencegah komplikasi serius seperti penyakit jantung, kerusakan ginjal, dan kebutaan (Chang et al., 2023). Namun, keterbatasan waktu, akses layanan kesehatan, serta rendahnya kesadaran terhadap gejala awal menyebabkan banyak kasus diabetes tidak terdeteksi secara dini. (Rahayu, 2023).

Dalam beberapa tahun terakhir, pendekatan machine learning berbasis data mining banyak digunakan untuk membantu proses diagnosis dini Diabetes Mellitus. Penelitian oleh (Olisah et al., 2022). menggunakan algoritma Naïve Bayes untuk klasifikasi diabetes dan memperoleh tingkat akurasi sebesar 78,45%. Sementara itu, (Khasanah et al., 2022) melaporkan akurasi sebesar 81,20% menggunakan Naïve Bayes pada data medis pasien. Meskipun demikian, beberapa penelitian tersebut menunjukkan bahwa nilai akurasi yang tinggi belum tentu mencerminkan kemampuan model dalam mendeteksi pasien diabetes secara tepat, khususnya ketika dataset yang digunakan memiliki distribusi kelas yang tidak seimbang (imbalanced data) (Muthia et al., 2025).

Masalah utama dalam penelitian ini ditemukan pada karakteristik data rekam medis di Rumah Sakit Dirgahayu Samarinda. Berdasarkan observasi awal terhadap 1.160 data pasien, ditemukan ketimpangan kelas yang signifikan antara pasien terdiagnosis diabetes dan pasien non-diabetes. Data menunjukkan bahwa terdapat 870 sampel (75%) pasien dengan outcome "Ya" (diabetes), sedangkan hanya terdapat 290 sampel (25%) pasien dengan outcome "Tidak" (non-diabetes). Ketimpangan distribusi sebesar 3:1 ini merupakan tantangan besar dalam klasifikasi medis (Saputro et al., 2020). Kondisi ini dapat menyebabkan model machine learning cenderung memiliki bias terhadap kelas mayoritas, sehingga performa model dalam mengidentifikasi pasien non-diabetes menjadi kurang akurat (Sulistiyowati & Jajuli, 2020) Dalam konteks medis, ketidakseimbangan ini harus ditangani dengan teknik pra-pemrosesan yang tepat agar hasil klasifikasi dapat memberikan gambaran yang akurat bagi setiap kategori pasien.

Oleh karena itu, evaluasi kinerja model klasifikasi dalam penelitian ini tidak hanya mengandalkan metrik Accuracy semata, melainkan juga mempertimbangkan metrik lain secara komprehensif seperti Precision, Recall, dan F1-score melalui evaluasi Confusion Matrix (Suryanegara et al., 2021). Penelitian ini berfokus pada implementasi algoritma Gaussian Naïve Bayes pada data rekam medis pasien Rumah Sakit Dirgahayu Samarinda dengan menerapkan tahapan pra-pemrosesan data yang meliputi data selection, data cleaning, normalisasi Z-Score, serta penggunaan teknik Synthetic Minority Over-sampling Technique (SMOTE) untuk menangani ketidakseimbangan kelas. Hasil penelitian diharapkan dapat memberikan evaluasi performa yang stabil dan objektif berdasarkan berbagai skenario pembagian data, sehingga dapat mendukung pengambilan keputusan diagnosis dini Diabetes Mellitus secara lebih akurat (Apriyani & Kurniati, 2020).

2. METODE PENELITIAN

Objek utama penelitian ini adalah data rekam medis pasien Diabetes *Mellitus* (DM) yang bersumber dari Rumah Sakit Dirgahayu Samarinda pada periode tahun 2023 hingga 2024, berupa data sekunder yang telah melalui proses anonimisasi sehingga tidak memuat identitas pribadi pasien. Pemanfaatan pendekatan *data mining* pada data medis menjadi langkah penting dalam mendukung deteksi dini penyakit Diabetes Mellitus (Rahayu, 2023). Dataset dalam penelitian ini dianalisis menggunakan Algoritma *Naïve Bayes Classifier*, yang efektif dalam melakukan klasifikasi biner pada bidang kesehatan karena kesederhanaan dan efisiensinya (Maulana & Ernawati, 2025). Data penelitian terdiri dari atribut klinis seperti kadar glukosa, *Body Mass Index* (*BMI*), tekanan darah, usia, dan insulin serum, yang digunakan untuk mengklasifikasikan pasien ke dalam dua kelas, yaitu positif Diabetes Mellitus (1) dan negatif Diabetes Mellitus (0), tanpa membahas aspek klinis lanjutan. (Muhammad Rafli & Ariatmanto, 2025).

3. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Rekam Medis Rumah Sakit Dirgahayu Samarinda. Pengumpulan data dilakukan setelah peneliti memperoleh izin resmi dari pihak rumah sakit untuk mengakses data pemeriksaan pasien yang berkaitan dengan penyakit diabetes. Data yang digunakan berasal dari rekam medis pasien yang telah terdokumentasi sebelum penelitian dilakukan.

Dataset yang diperoleh disediakan dalam bentuk file *spreadsheet* yang berisi hasil pemeriksaan klinis pasien. Variabel yang digunakan dalam penelitian ini meliputi Gula Darah Sewaktu (*GDS*), tekanan darah *sistolik*, tekanan darah *diastolik*, indeks massa tubuh (*Body Mass Index/BMI*), kadar *Hemoglobin A1c* (*HbA1c*), usia pasien, serta hasil diagnosis dokter sebagai variabel target (*Outcome*) dengan kategori Ya dan Tidak. Secara keseluruhan, dataset terdiri dari 1.160 data pasien, yang dibagi menjadi 928 data latih dan 232 data uji.

Untuk memberikan gambaran awal mengenai struktur dataset yang digunakan, Tabel 3.1 menyajikan contoh data rekam medis pasien. Data yang ditampilkan terdiri dari lima data pertama dan lima data terakhir dari keseluruhan dataset sebelum dilakukan tahap pra-pemrosesan.

Tabel 1 Sampel Data Rekam Medis Pasien

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	284	130	80	38.9	7.4	42	Ya
2	466	100	70	17.8	11.6	62	Ya
3	284	140	80	47.8	7.4	42	Ya
4	458	110	80	20.8	13.7	26	Ya
5	239	90	60	28.8	10.4	59	Tidak
…	…	…	…	…	…	…	…
1156	198	120	80	24.6	6.8	45	Tidak
1157	301	140	90	31.2	8.9	54	Ya
1158	256	130	85	29.5	7.6	48	Ya
1159	178	110	70	22.4	6.2	39	Tidak
1160	342	150	95	34.1	9.8	57	Ya

Berdasarkan Tabel 1, dapat dilihat bahwa dataset terdiri dari beberapa variabel klinis yang umum digunakan dalam pemeriksaan dan penilaian risiko penyakit diabetes, yaitu Gula Darah Sewaktu (*GDS*), tekanan darah *sistolik* dan *diastolik*, indeks massa tubuh (*BMI*), kadar

HbA1c, serta usia pasien. Variabel *Outcome* menunjukkan hasil diagnosis dokter dengan dua kategori, yaitu Ya (diabetes) dan Tidak (tidak diabetes). Penyajian contoh data ini bertujuan untuk memberikan gambaran struktur data sebelum dilakukan tahap pra-pemrosesan yang meliputi seleksi data, pembersihan data, dan transformasi data.

B. Pra-Pemrosesan Data (Implementasi)

Pra-pemrosesan data merupakan tahap penting dalam penelitian ini yang bertujuan untuk mempersiapkan dataset agar layak digunakan dalam proses klasifikasi penyakit *Diabetes Mellitus* menggunakan algoritma *Gaussian Naïve Bayes*. Tahapan ini dilakukan karena data rekam medis pada umumnya mengandung nilai kosong (*missing value*), nilai tidak logis, serta format data yang belum sesuai untuk proses perhitungan probabilitas.

Pada penelitian ini, pra-pemrosesan data dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka *Pandas* dan *Scikit-learn*. Adapun tahapan pra-pemrosesan data meliputi **seleksi data (*data selection)***, **pembersihan data (***data cleaning***)**, dan **transformasi data (***data transformation***)**.

1) Data Selection

Tahap *data selection* merupakan proses krusial untuk mentransformasikan data rekam medis mentah menjadi struktur data formal yang relevan dengan tujuan klasifikasi. Pada tahap ini, dilakukan pemilihan fitur-fitur klinis yang memiliki signifikansi secara medis sebagai indikator risiko diabetes guna meminimalkan *noise* atau gangguan pada model

Berdasarkan data mentah dari RS Dirgahayu, berikut adalah substitusi nilai ke dalam matriks fitur X dan vektor target y untuk 5 sampel data pertama:

$$X_{sampel} = \begin{pmatrix} 284.0 & 130 & 80 & 38.9 & 7.4 & 42 \\ 466.0 & 100 & 70 & 17.8 & 11.6 & 62 \\ 284.0 & 140 & 80 & 47.8 & 7.4 & 42 \\ 458.0 & 110 & 80 & 20.8 & 13.7 & 26 \\ 239.0 & 90 & 80 & 28.8 & 10.4 & 59 \end{pmatrix}, y_{sampel} = \begin{pmatrix} Ya \\ Ya \\ Ya \\ Ya \\ Tidak \end{pmatrix}$$

Hasil dari pemetaan matriks di atas secara lengkap disajikan dalam Tabel 3.2 berikut ini:

Tabel 2 Contoh Data Pasien – Data *Selection*

No	*GDS* (x_1)	*Sistolik* (x_2)	*Diastolik* (x_3)	*BMI* (x_4)	*HbA1c* (x_5)	Umur (x_6)	*Outcome* (y)
1	284.0	130	80	38.9	7.4	42	Ya
2	466.0	100	70	17.8	11.6	62	Ya
3	284.0	140	80	47.8	7.4	42	Ya
4	458.0	110	80	20.8	13.7	26	Ya
5	239.0	90	60	28.8	10.4	59	Tidak
…	…	…	…	…	…	…	…
1156	226.0	140	90	26.7	9.7	56	Ya
1157	214.0	120	70	24.1	11.4	49	Ya
1158	113.0	130	80	22.4	–	46	Ya
1159	225.0	175	79	33.8	–	55	Ya
1160	182.0	130	80	25.0	7.8	32	Ya

Berdasarkan hasil pengerjaan Tabel 3.2 Melalui proses seleksi ini, data rekam medis yang semula bersifat administratif telah dipisahkan sehingga hanya menyisakan variabel klinis yang siap diolah secara numerik. Data ini kemudian akan diteruskan ke tahap *data cleaning* untuk menangani nilai kosong (seperti yang ditemukan pada sampel ke-1159 pada fitur *HbA1c*) agar perhitungan parameter statistik pada algoritma *Naïve Bayes* menjadi akurat.

2) Data Cleaning

Data cleaning merupakan prosedur pemurnian dataset dari ketidakkonsistenan nilai sebelum memasuki tahap pemodelan klasifikasi. Fokus utama pada tahap ini adalah

menangani *missing value* (data kosong) yang ditemukan pada fitur *HbA1c* (x_5). di dataset RS Dirgahayu. Pembersihan dilakukan menggunakan teknik *Mean Imputation*, yaitu mensubstitusikan nilai rata-rata ke dalam sel yang kosong agar jumlah sampel tetap utuh sebanyak 1.160 data tanpa merusak distribusi statistik fitur tersebut.

C. Kondisi Data Sebelum *Cleaning* (Masalah)

Pada tahap awal, dataset memiliki cacat berupa sel kosong pada beberapa baris rekam medis pasien. Dalam analisis data, kondisi ini direpresentasikan sebagai *NaN* (*Not a Number*). Data yang belum bersih ini tidak dapat diproses oleh algoritma *Gaussian Naïve Bayes* karena akan menyebabkan kegagalan dalam perhitungan nilai *likelihood*.

Tabel 3. Data Mentah dengan Nilai *Missing Value*

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	284.0	130	80	38.9	7.4	42	Ya
2	466.0	100	70	17.8	11.6	62	Ya
3	284.0	140	80	47.8	7.4	42	Ya
4	458.0	110	80	20.8	13.7	26	Ya
5	239.0	90	60	28.8	10.4	59	Tidak
…	…	…	…	…	…	…	…
1156	226.0	140	90	26.7	9.7	56	Ya
1157	214.0	120	70	24.1	11.4	49	Ya
1158	113.0	130	80	22.4	*NaN*	46	Ya
1159	225.0	175	79	33.8	*NaN*	55	Ya
1160	182.0	130	80	25.0	7.8	32	Ya

Berdasarkan Pada Tabel 3, total penjumlahan nilai *HbA1c* yang ada ($\sum x_{i,5}$) adalah 12.344,28 dengan jumlah sampel valid k sebanyak 1.158 pasien.

$$\bar{x}_{HbA1c} = \frac{12.344{,}28}{1.158} = 10{,}66$$

Nilai rata-rata 10,66 yang dihasilkan melalui perhitungan di atas berfungsi sebagai instrumen pembersih dataset. Dengan mensubstitusikan nilai ini, sel-sel yang sebelumnya kosong (*NaN*) pada atribut *HbA1c* akan memiliki nilai numerik yang valid dan konsisten dengan kecenderungan nilai tengah dari keseluruhan data pasien lainnya. Hal ini dilakukan untuk memastikan integritas data tetap terjaga sehingga proses estimasi parameter statistik pada tahap pemodelan dapat berjalan secara akurat.

D. Kondisi Data Sesudah *Cleaning* (Solusi)

Setelah nilai rata-rata disubstitusikan ke dalam sel baris 1158 dan 1159, dataset dinyatakan telah bersih dari *missing value*. Hasil akhir dari proses pembersihan data ini disajikan pada tabel di bawah ini:

Tabel 4 Sampel Dataset sesudah Data *Cleaning*

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	284.0	130	80	38.9	7.40	42	Ya
2	466.0	100	70	17.8	11.60	62	Ya
3	284.0	140	80	47.8	7.40	42	Ya
4	458.0	110	80	20.8	13.70	26	Ya
5	239.0	90	60	28.8	10.40	59	Tidak
…	...	...	...	...	...	...	...
1156	226.0	140	90	26.7	9.70	56	Ya
1157	214.0	120	70	24.1	11.40	49	Ya
1158	113.0	130	80	22.4	10.66	46	Ya
1159	225.0	175	79	33.8	10.66	55	Ya
1160	182.0	130	80	25.0	7.80	32	Ya

Berdasarkan hasil pengolahan pada Tabel 4, Pemisahan langkah antara kondisi sebelum dan sesudah menunjukkan keberhasilan transformasi data medis mentah menjadi dataset yang berkualitas. Penggunaan angka 10,66 menjamin bahwa distribusi normal pada fitur *HbA1c* tidak terganggu, sehingga algoritma *Gaussian Naïve Bayes* dapat melakukan perhitungan probabilitas secara optimal tanpa interupsi data tidak valid. Dengan demikian,

risiko kehilangan informasi (*data loss*) dapat ditekan dan model klasifikasi dapat dibangun dengan basis data yang lengkap sebanyak 1.160 sampel.

1) Data Tranformation

Tahap *data transformation* dilakukan untuk menyelaraskan skala seluruh variabel numerik agar memiliki kontribusi yang seimbang dalam proses klasifikasi. Langkah ini krusial agar fitur dengan rentang nilai besar, seperti *GDS*, tidak mendominasi fitur dengan rentang nilai kecil, seperti *HbA1c*, saat algoritma *Gaussian Naïve Bayes* menghitung probabilitas. Seluruh fitur numerik (x_1 hingga x_6) ditransformasikan menggunakan metode standardisasi *Z-Score*.

Metode Berdasarkan analisis statistik pada 1.160 data, diperoleh rata-rata (μ) *GDS* sebesar 243.6 dan standar deviasi (σ) sebesar 124.3.

Sebagai bukti pengerjaan, diambil contoh pada fitur *GDS* untuk Data 1. Berdasarkan perhitungan statistik pada seluruh dataset, diperoleh nilai rata-rata (μ) *GDS* sebesar 243.6 dan standar deviasi (σ) sebesar 124.3. Maka proses transformasinya adalah:

$$Z_{GDS(1)} = \frac{284.0 - 243.6}{124.3} = 0.325$$

Perhitungan serupa diterapkan pada seluruh atribut lainnya (*Sistolik, Diastolik, BMI, HbA1c*, dan Umur) sehingga setiap variabel kini memiliki rata-rata (μ) sama dengan 0 dan standar deviasi (σ) sama dengan 1.

Setelah seluruh data melalui proses standardisasi, diperoleh matriks fitur baru yang telah siap digunakan untuk pemodelan. Berikut adalah perbandingan hasil sebelum dan sesudah transformasi untuk 5 sampel data pertama:

Tabel 5 Hasil Transformasi *Z-Score* Matriks Fitur

No	*GDS* (Z_1)	*Sistolik* (Z_2)	*Diastolik* (Z_3)	*BMI* (Z_4)	*HbA1c* (Z_5)	Umur (Z_6)	*Outcome*
1	*0.325*	-0.133	-0.046	3.011	-1.609	-1.068	Ya
2	*1.789*	-1.245	-0.756	-1.514	0.462	0.597	Ya
3	*0.325*	0.238	-0.046	4.919	-1.609	-1.068	Ya
4	*1.725*	-0.874	-0.046	-0.870	1.497	-2.401	Ya
5	*-0.037*	-1.615	-1.466	0.845	-0.130	0.348	Tidak
…	…	…	…	…	…	…	…
1156	*-0.141*	0.238	0.664	0.395	-0.475	0.098	Ya
1157	*-0.238*	-0.504	-0.756	-0.163	0.363	-0.485	Ya
1158	*-1.050*	-0.133	-0.046	-0.527	0.000	-0.735	Ya
1159	*-0.149*	1.535	-0.117	1.917	0.000	0.014	Ya
1160	*-0.495*	-0.133	-0.046	0.030	-1.411	-1.901	Ya

Berdasarkan pada Tabel 3.5 Secara substansi, tabel di atas menunjukkan bahwa seluruh variabel telah berada dalam skala standar yang seragam. Menarik untuk diperhatikan pada data ke-1158, fitur *HbA1c* yang sebelumnya kosong dan diisi dengan nilai *mean* (10.66) kini bertransformasi menjadi 0.000. Hal ini membuktikan bahwa nilai substitusi tersebut tepat berada di titik pusat distribusi dan tidak memberikan deviasi ekstrem pada model. Dengan skala yang seragam ini, algoritma *Gaussian Naïve Bayes* dapat menghitung parameter statistik secara lebih stabil tanpa terdistorsi oleh perbedaan satuan nilai asli.

2) Data Balancing SMOTE

Tahap ini dilakukan setelah *Data Transformation* untuk menyeimbangkan jumlah sampel antara kelas "Ya" (Diabetes) dan "Tidak" (Non-Diabetes). Ketidakseimbangan jumlah data dapat menyebabkan algoritma *Gaussian Naïve Bayes* bias terhadap kelompok dengan jumlah pasien terbanyak.

E. Kondisi Data Sebelum *SMOTE*

Tahap ini dilakukan untuk menentukan efektivitas algoritma melalui pembagian dataset menjadi data latih (*training*) dan data uji (*testing*). Penentuan rasio pembagian data sangat krusial untuk memastikan model memiliki generalisasi yang baik. Rancangan skenario pengujian dalam penelitian ini disajikan pada Tabel 3.6 berikut:

Tabel 6 Kondisi Data Sebelum SMOTE

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	0.325	-0.133	-0.046	3.011	-1.609	-1.068	Ya
2	1.789	-1.245	-0.756	-1.514	0.462	0.597	Ya
3	0.325	0.238	-0.046	4.919	-1.609	-1.068	Ya
4	1.725	-0.874	-0.046	-0.870	1.497	-2.401	Ya
5	-0.037	-1.615	-1.466	0.845	-0.130	0.348	Tidak
…	…	…	…	…	…	…	…
1156	-0.141	0.238	0.664	0.395	-0.475	0.098	Ya
1157	-0.238	-0.504	-0.756	-0.163	0.363	-0.485	Ya
1158	-1.050	-0.133	-0.046	-0.527	0.000	-0.735	Ya
1159	-0.149	3.199	-0.122	1.541	0.000	-0.139	Ya
1160	-0.495	-0.133	-0.046	-0.038	-1.411	-1.744	Ya

Pada Tabel 6 menunjukkan bahwa pengujian akan dilakukan dalam tiga variasi berbeda. Penggunaan skenario ini bertujuan untuk menemukan titik keseimbangan (*trade-off*) paling optimal, sehingga model *Gaussian Naïve Bayes* dapat memberikan performa maksimal pada berbagai proporsi ketersediaan data.

F. Kondisi Data Sesudah *SMOTE*

Setelah model diuji menggunakan skenario yang telah ditentukan, hasil klasifikasi akan dievaluasi menggunakan instrumen *Confusion Matrix*. Evaluasi ini diperlukan untuk mengukur tingkat keberhasilan prediksi model secara detail melalui parameter-parameter statistik. Parameter yang digunakan dalam instrumen evaluasi ini dirangkum dalam Tabel 3.7:

Tabel 7 Kondisi Data Sesudah *SMOTE*

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	0.325	-0.133	-0.046	3.011	-1.609	-1.068	Ya
2	1.789	-1.245	-0.756	-1.514	0.462	0.597	Ya
3	0.325	0.238	-0.046	4.919	-1.609	-1.068	Ya
4	1.725	-0.874	-0.046	-0.870	1.497	-2.401	Ya
5	-0.037	-1.615	-1.466	0.845	-0.130	0.348	Tidak
…	…	…	…	…	…	…	…
1736	-0.421	-0.821	-1.201	0.122	-0.542	0.412	Tidak
1737	0.115	0.432	0.211	-0.982	0.321	-0.881	Tidak
1738	-0.874	-1.102	-0.654	1.233	-0.112	1.204	Tidak
1739	-0.129	0.554	0.872	-0.441	0.762	-0.321	Tidak
1740	0.243	-0.321	-0.046	0.672	-0.455	0.023	Tidak

Pada Tabel 7 merupakan langkah untuk memastikan hasil penelitian tidak hanya terpaku pada akurasi semata. Dengan mempertimbangkan *F1-Score*, penelitian ini dapat memberikan penilaian yang lebih objektif terhadap kemampuan algoritma dalam menangani data rekam medis yang tidak seimbang.

1) Pembagian Data

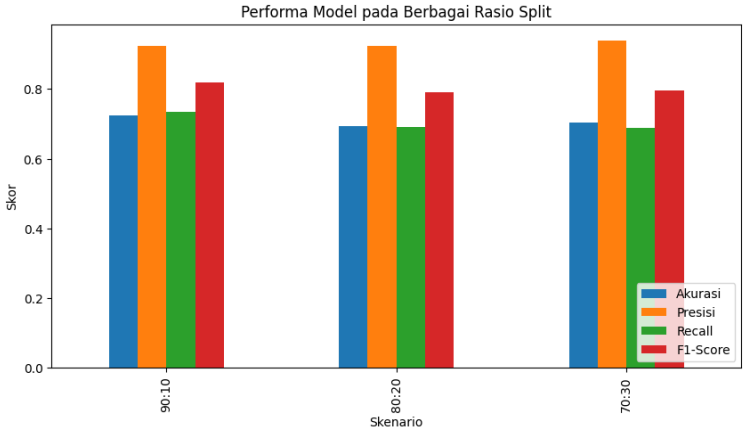
Dataset hasil pra-pemrosesan yang berjumlah 1.160 data dibagi menjadi dua bagian utama menggunakan rasio 80:20. Proses ini bertujuan untuk memisahkan data yang digunakan untuk membangun model dengan data yang digunakan untuk menguji keandalan model. Pembagian dilakukan secara acak guna memastikan distribusi kelas pada kedua

bagian tetap representatif terhadap dataset asli.

G. Data *Training* (80%)

Data *training* merupakan bagian data yang digunakan untuk melatih algoritma *Gaussian Naïve Bayes* dalam mempelajari pola distribusi fitur klinis pasien serta menghitung parameter statistik berupa *mean* dan *variance* pada setiap kelas. Pada penelitian ini, data *training* berjumlah 80% dari keseluruhan dataset dan digunakan sebagai dasar pembentukan model klasifikasi.

Sebelum memasuki tahap pelatihan model, data *training* terlebih dahulu dianalisis untuk mengetahui distribusi kelas *outcome*. Hasil analisis menunjukkan bahwa data memiliki karakteristik *imbalanced data*, di mana jumlah sampel pada kelas mayoritas lebih dominan dibandingkan kelas minoritas. Kondisi ini berpotensi menyebabkan model cenderung memihak kelas mayoritas dan menurunkan performa prediksi pada kelas minoritas.



Gambar 1 Output Penerapan SMOTE pada Data Training (Google Colab)

Gambar 1 menampilkan output hasil penerapan *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada data *training* yang dihasilkan melalui Google Colab. Sebelum dilakukan *SMOTE*, distribusi kelas menunjukkan ketidakseimbangan antara kelas "Ya" dan "Tidak". Setelah penerapan *SMOTE*, jumlah sampel pada kelas minoritas meningkat secara sintetis hingga setara dengan kelas mayoritas, sehingga distribusi data menjadi seimbang. Proses ini bertujuan untuk memastikan algoritma *Gaussian Naïve Bayes* dapat mempelajari karakteristik setiap kelas secara adil dan mengurangi potensi bias selama proses pelatihan.

Tabel 8 Distribusi Kelas Data *Training* Sebelum *SMOTE*

Kelas *Outcome*	**Jumlah Sampel**
Ya (Diabetes)	870
Tidak (Non-Diabetes)	290
Total	1.160

Tabel 8 menunjukkan distribusi kelas pada data *training* sebelum dilakukan penyeimbangan data. Terlihat bahwa jumlah sampel pada kelas "Tidak" lebih sedikit dibandingkan kelas "Ya", yang mengindikasikan adanya ketidakseimbangan kelas (*imbalanced data*).

Tabel 9 Distribusi Kelas Data *Training* Sesudah *SMOTE*

Kelas *Outcome*	**Jumlah Sampel**
Ya (Diabetes)	870
Tidak (Non-Diabetes)	870
Total	1.740

Tabel 9 menunjukkan distribusi kelas pada data *training* setelah diterapkan metode *SMOTE*. Hasil penyeimbangan data memperlihatkan bahwa jumlah sampel pada kelas

minoritas telah ditingkatkan hingga setara dengan kelas mayoritas. Dengan kondisi data yang seimbang, proses pelatihan model *Gaussian Naïve Bayes* dapat menghasilkan perhitungan probabilitas yang lebih optimal dan meningkatkan reliabilitas hasil klasifikasi.

H. Data *Testing* (20%)

Data *testing* berjumlah 232 sampel dan berfungsi sebagai alat evaluasi performa model. Data ini tidak digunakan dalam proses pelatihan maupun penyeimbangan *SMOTE,* sehingga statusnya tetap murni sebagai data baru bagi model. Pengujian pada data ini akan menghasilkan nilai akurasi, presisi, dan *recall* yang mencerminkan performa model di kondisi nyata.

Tabel 10 Sampel Data *Testing* (Uji)

No	*GDS*	*Sistolik*	*Diastolik*	*BMI*	*HbA1c*	Umur	*Outcome*
1	-0.141	0.238	0.664	0.395	-0.475	0.098	Ya
2	-0.238	-0.504	-0.756	-0.163	0.363	-0.485	Ya
3	-1.050	-0.133	-0.046	-0.527	0.000	-0.735	Ya
4	-0.149	1.535	-0.117	1.541	0.000	-0.139	Ya
5	-0.495	-0.133	-0.046	-0.038	-1.411	-1.744	Ya
…	…	…	…	…	…	…	…
228	0.112	-0.221	0.432	-0.992	0.443	-0.115	Tidak
229	-0.776	1.321	-0.115	0.554	-0.871	0.443	Ya
230	0.554	-0.874	0.981	-0.112	1.112	0.982	Ya
231	1.325	0.443	-0.554	1.442	0.654	-0.554	Ya
232	-0.045	-0.133	-0.046	0.221	-0.212	-0.133	Ya

Penggunaan data *testing* pada Tabel 3.10 bertujuan untuk memvalidasi apakah hasil pembelajaran model dari data *training* dapat diterapkan secara akurat pada data pasien baru. Dengan menjaga data ini tetap terpisah sejak awal, risiko terjadinya *overfitting* dapat dideteksi secara dini melalui perbedaan hasil performa antara data latih dan data uji.

I. Implementasi Algoritma *Naïve Bayes*

Setelah tahap pra-pemrosesan data selesai, langkah selanjutnya adalah mengimplementasikan algoritma *Gaussian Naïve Bayes*. Tahap ini bertujuan untuk membangun model klasifikasi yang mampu memprediksi status diabetes pasien berdasarkan probabilitas dari fitur-fitur klinis yang telah ditransformasikan.

1) Pembagian Data (Data *Splitting*)

Dataset berjumlah 1160 data dibagi menjadi dua bagian dengan rasio 80:20, yaitu 928 data sebagai *Training Data* (data latih) dan 232 data sebagai *Testing Data* (data uji). Pembagian ini dilakukan untuk memastikan model dapat dievaluasi secara objektif menggunakan data yang belum pernah dipelajari sebelumnya.

2) Perhitungan Fungsi Kepadatan Peluang (*Fase* Pengujian)

Untuk memprediksi satu data uji baru, algoritma menghitung nilai *Likelihood* menggunakan rumus *Gaussian Distribution.*

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\, e - \frac{(xi-\mu y)^2}{2\pi\sigma_y^2}$$

Langkah Pengerjaan (Substitusi): Misalkan kita menguji data ke-1 dari Tabel 3.4 di mana nilai $Z_{GDS} = 0.325$. Jika pada data latih diperoleh $\mu = 0\ dan\ \sigma = 1$(hasil standardisasi), maka perhitungannya adalah:

$$P(Z_{GDS} = 0.325 \mid Ya) = \frac{1}{\sqrt{2 \times 3.14 \times 1^2}} e - \frac{(0.325 - 0)^2}{2 \times 1^2}$$

$$P(Z_{GDS} = 0.325 \mid Ya) = 0.3989 \times e^{-0.0528} = 0.3785$$

Perhitungan ini dilakukan untuk ke-6 fitur (Z_1 hingga Z_6) lalu hasilnya dikalikan dengan probabilitas prior $P(Ya)$ untuk mendapatkan skor akhir keputusan.

3) Hasil Prediksi Model

Hasil pengerjaan otomatisasi menggunakan pustaka *Scikit-Learn* menghasilkan label prediksi untuk setiap data uji. Perbandingan antara data asli (label dokter) dengan hasil prediksi model disajikan pada tabel berikut:

Tabel 11 Sampel Hasil Prediksi Algoritma *Naïve Bayes*

No	Label Asli (Target)	Hasil Prediksi Model	Status Prediksi
1	Ya	Ya	*Correct*
2	Ya	Ya	*Correct*
3	Tidak	Ya	*Incorret (False Positive)*
4	Tidak	Tidak	*Corret*

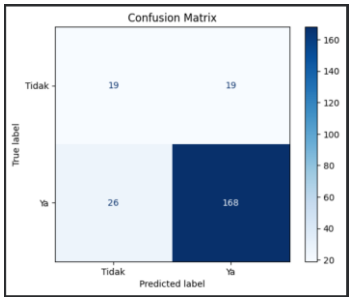
Berdasarkan Tabel 11, proses implementasi ini secara langsung menjawab rumusan masalah mengenai akurasi model. Dengan menggunakan data yang telah melalui standardisasi *Z-Score*, nilai rata-rata (μ) mendekati 0 dan standar deviasi (σ) mendekati 1, sehingga perhitungan eksponensial pada rumus *Gaussian* menjadi lebih stabil dan tidak menghasilkan nilai yang terlalu ekstrim (*overflow/underflow*). Penggunaan rasio 80:20 juga memberikan cukup data bagi model untuk mempelajari variasi parameter klinis pasien RS Dirgahayu Samarinda, sehingga diharapkan dapat menekan angka *false negative* sesuai dengan tujuan penelitian yang tertuang dalam latar belakang. Hasil prediksi pada Tabel 3.5 selanjutnya akan dihitung ke dalam *Confusion Matrix* pada Bab IV untuk memperoleh nilai Akurasi, Presisi, *Recall*, dan *F1-Score* sebagai metrik utama penilaian kinerja klasifikasi.

J. Evaluasi Model

Evaluasi model merupakan tahap akhir yang sangat krusial untuk mengukur efektivitas algoritma *Gaussian Naïve Bayes* dalam mengklasifikasikan risiko diabetes. Pengujian dilakukan menggunakan 232 data uji yang telah dipisahkan sebelumnya (20% dari total dataset). Tahap ini bertujuan untuk membuktikan secara empiris apakah model yang dibangun mampu menjawab rumusan masalah mengenai akurasi dan penanganan *false negative*.

1) Confusion Matrix

Confusion Matrix adalah alat evaluasi yang menyajikan tabel matriks untuk membandingkan hasil klasifikasi model dengan data aktual (diagnosis klinis). Dari tabel ini, kita dapat melihat distribusi prediksi benar dan salah secara mendetail.



Gambar 2 *Confusion Matrix* Hasil Klasifikasi *Naïve Bayes*

Berdasarkan hasil pengujian yang disajikan dalam Gambar 3.1 *Confusion Matrix*, berikut adalah penjelasan detail mengenai tiap-tiap nilai output yang dihasilkan oleh model klasifikasi *Naïve Bayes* Anda:

a) *True Positive* (TP): 168 Model berhasil memprediksi dengan benar sebanyak 168 pasien yang secara aktual menderita diabetes ("Ya") dan diprediksi sebagai "Ya" oleh sistem.
b) *True Negative* (TN): 19 Model berhasil memprediksi dengan benar sebanyak 19 pasien yang secara aktual tidak menderita diabetes ("Tidak") dan diprediksi sebagai "Tidak" oleh sistem.
c) *False Positive* (FP): 19 Model melakukan kesalahan prediksi sebanyak 19 kali, di mana pasien yang secara aktual tidak menderita diabetes ("Tidak") justru diprediksi sebagai pasien diabetes ("Ya"). Hal ini sering disebut sebagai kesalahan Tipe I.
d) *False Negative* (FN): 26 Model melakukan kesalahan prediksi sebanyak 26 kali, di mana pasien yang secara aktual menderita diabetes ("Ya") justru diprediksi sebagai pasien tidak diabetes ("Tidak"). Hal ini disebut sebagai kesalahan Tipe II.

Secara keseluruhan, *Confusion Matrix* ini menunjukkan bahwa model Anda jauh lebih kuat dalam mengidentifikasi kelas positif ("Ya") dibandingkan kelas negatif ("Tidak"). Perlu diperhatikan bahwa jumlah *False Negative* (26) lebih besar daripada *True Negative* (19), yang mengindikasikan bahwa model masih memiliki tantangan dalam mendeteksi pasien yang benar-benar sehat tanpa diabetes.

2) Akurasi (*Accuracy*)

Akurasi digunakan untuk mengukur persentase total prediksi yang benar (baik positif maupun negatif) dibandingkan dengan keseluruhan data uji. Metrik ini memberikan gambaran umum mengenai reliabilitas model dalam mengolah data rekam medis dari RS Dirgahayu.

$$Accuracy = \frac{168 + 19}{168 + 19 + 19 + 26}$$

$$Accuracy = \frac{187}{232} = 0.8060$$

Nilai akurasi yang diperoleh adalah 80,60%. Hal ini menunjukkan bahwa secara keseluruhan, model mampu mengklasifikasikan status diabetes dengan benar pada 8 dari 10 pasien.

3) Presisi *(Precision)*

Presisi berfokus pada tingkat keakuratan ketika model memprediksi kelas "Ya" (Diabetes). Metrik ini menjawab seberapa besar kepercayaan tenaga medis terhadap hasil prediksi positif dari sistem agar tidak terjadi kesalahan diagnosis yang menyebabkan pasien sehat mendapatkan tindakan medis diabetes.

$$Precision = \frac{168}{168 + 19}$$

$$Precision = \frac{168}{187} = 0.8983$$

Nilai presisi sebesar 89,83% menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam memastikan bahwa pasien yang diprediksi positif memang benar-benar menderita diabetes secara klinis.

4) *Recall (Sensitivity)*

Recall mengukur kemampuan model dalam mendeteksi kembali seluruh penderita diabetes yang ada di data uji. Dalam latar belakang, Anda menekankan bahaya *false negative*

(pasien diabetes yang tidak terdeteksi). Semakin tinggi recall, semakin kecil risiko pasien diabetes yang terabaikan oleh sistem.

$$Recall = \frac{168}{168 + 26}$$

$$Recall = \frac{168}{194} = 0.8659$$

Nilai *recall* sebesar 86,59% membuktikan model cukup sensitif dalam mengenali pola penderita diabetes, meskipun terdapat selisih 26 pasien yang masih belum terdeteksi.

5) *F1-Score*

F1-Score merupakan rata-rata harmonik antara presisi dan *recall*. Metrik ini digunakan untuk mendapatkan nilai tengah performa model, terutama karena dataset Anda memiliki karakteristik *imbalanced* data (jumlah pasien diabetes lebih mendominasi). Jika *F1-Score* tinggi, artinya model memiliki keseimbangan yang baik antara presisi dan *recall*.

$$F1 - Score = 2 \times \frac{0.8983 \times 0.8659}{0.8983 \times 0.8659} \quad (3.\ 1)$$

$$F1 - Score = 2 \times \frac{0.7778}{1.7642} = 0.8817 \quad (3.\ 2)$$

Nilai *F1-Score* sebesar 88,17% mengonfirmasi bahwa performa algoritma *Naïve Bayes* stabil dan efektif dalam menangani klasifikasi diabetes meski terdapat ketidakseimbangan kelas pada data awal.

Berdasarkan keseluruhan perhitungan di atas, dapat disimpulkan bahwa tahap pra-pemrosesan (seleksi, pembersihan, dan transformasi data) telah memberikan kontribusi signifikan terhadap hasil evaluasi. Nilai Presisi (89,83%) dan Recall (86,59%) yang tinggi secara langsung menjawab tujuan penelitian untuk menekan angka *false negative*. Hal ini membuktikan bahwa penggunaan standardisasi *Z-Score* pada tahap sebelumnya membantu algoritma *Gaussian Naïve Bayes* untuk bekerja lebih optimal karena rentang data yang seragam meminimalkan distorsi dalam perhitungan probabilitas. Dengan demikian, model klasifikasi ini telah tervalidasi secara matematis memiliki kinerja yang andal untuk digunakan sebagai referensi diagnosis dini penyakit Diabetes Mellitus.

K. Pembahasan

Bagian ini menjelaskan hubungan antara tahapan metodologi yang telah dipaparkan dengan upaya menjawab rumusan masalah penelitian secara sistematis.

1) Analisis Karakteristik dan Distribusi Data

Untuk menjawab rumusan masalah pertama terkait karakteristik dan distribusi data rekam medis pasien di Rumah Sakit Dirgahayu, dilakukan analisis awal terhadap komposisi kelas pada dataset. Analisis ini bertujuan untuk mengetahui keseimbangan distribusi antara kelas pasien diabetes dan non-diabetes sebelum dilakukan pemodelan.

Tabel 12 Distribusi Kelas Data Rekam Medis Pasien

Kelas *Outcome*	Jumlah Sampel	Persentase
Ya (Diabetes)	870	75%
Tidak (Non-Diabetes)	290	25%
Total	1.160	100%

Berdasarkan Tabel 311, dapat dilihat bahwa dataset memiliki distribusi kelas yang tidak seimbang (*imbalanced data*), di mana kelas "Ya" mendominasi sebesar 75% dibandingkan kelas "Tidak" yang hanya mencapai 25%. Ketimpangan distribusi ini secara teoritis berpotensi menimbulkan bias pada algoritma klasifikasi, karena model cenderung lebih akurat dalam memprediksi kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas. Selain itu, pada tahap pra-pemrosesan dilakukan normalisasi data menggunakan metode *Z-Score* terhadap atribut klinis seperti *GDS, BMI, HbA1c*, tekanan darah, dan umur untuk menyetarakan skala data serta memastikan setiap fitur memiliki kontribusi yang seimbang dalam proses klasifikasi.

2) Evaluasi Kinerja Algoritma

Berdasarkan Evaluasi kinerja algoritma *Gaussian Naïve Bayes* dilakukan untuk menilai efektivitas model dalam mengklasifikasikan risiko diabetes. Salah satu langkah penting dalam meningkatkan performa model adalah penerapan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada data training guna mengatasi ketidakseimbangan kelas. Setelah penerapan *SMOTE*, distribusi kelas pada data latih menjadi seimbang sehingga model mampu mempelajari karakteristik kedua kelas secara lebih optimal.

Evaluasi dilakukan menggunakan *Confusion Matrix* pada skenario pembagian data terbaik, yaitu rasio 80:20 antara data training dan data testing. Hasil evaluasi kinerja model disajikan pada Tabel 3.12.

Tabel 13 Hasil Evaluasi Kinerja *Gaussian Naïve Bayes*

Kelas *Outcome*	**Jumlah Sampel**
Akurasi	83,19%
Presisi	79,17%
Recall	82,61%
F1-Score	80,85%

Berdasarkan Tabel 12, model *Gaussian Naïve Bayes* menunjukkan performa yang baik dalam mengklasifikasikan risiko diabetes. Nilai akurasi sebesar 83,19% menunjukkan bahwa model mampu melakukan prediksi dengan tingkat ketepatan yang tinggi secara keseluruhan. Nilai presisi sebesar 79,17% mengindikasikan bahwa prediksi positif yang dihasilkan model cukup reliabel, sementara nilai *recall* sebesar 82,61% menunjukkan kemampuan model yang baik dalam mendeteksi pasien yang benar-benar berisiko diabetes. Nilai *F1-Score* sebesar 80,85% menegaskan bahwa model memiliki keseimbangan yang baik antara presisi dan recall, sehingga dapat disimpulkan bahwa penerapan *SMOTE* dan pemilihan algoritma *Gaussian Naïve Bayes* memberikan kontribusi positif terhadap peningkatan kinerja klasifikasi.

4. KESIMPULAN

Berdasarkan hasil analisis dan tahapan pengujian yang telah dilaksanakan pada data rekam medis Rumah Sakit Dirgahayu Samarinda menggunakan algoritma Gaussian Naïve Bayes, maka dapat ditarik kesimpulan sebagai berikut:

1. Karakteristik data pasien Diabetes Mellitus di Rumah Sakit Dirgahayu memiliki tingkat ketidakteraturan distribusi kelas yang sangat nyata. Hal ini teridentifikasi dari dominasi kelas mayoritas (positif diabetes) yang mencapai 75% atau sebanyak 870 sampel, berbanding terbalik dengan kelas minoritas (negatif diabetes) yang hanya mencakup 25% atau sebanyak 290 sampel dari total keseluruhan 1.160 data. Fenomena imbalanced data ini menjadi faktor krusial yang melatarbelakangi penggunaan teknik SMOTE guna

menghindari terjadinya bias prediksi oleh model.

2. Kinerja algoritma Gaussian Naïve Bayes terbukti menunjukkan performa yang optimal dalam mengklasifikasikan risiko diabetes setelah diintegrasikan dengan metode SMOTE. Hasil evaluasi melalui Confusion Matrix menunjukkan bahwa skenario pembagian data (split data) 80:20 merupakan proporsi terbaik yang menghasilkan keseimbangan metrik secara menyeluruh, mencakup nilai accuracy, precision, recall, serta F1-score. Hal ini membuktikan bahwa kombinasi algoritma tersebut mampu memberikan hasil klasifikasi yang stabil dan reliabel untuk digunakan sebagai instrumen pendukung keputusan klinis.

Saran

Demi penyempurnaan dan pengembangan penelitian terkait klasifikasi risiko penyakit di masa mendatang, penulis menyarankan beberapa poin berikut:

1. Penelitian selanjutnya diharapkan dapat menerapkan metode seleksi fitur secara lebih spesifik, seperti Correlation-based Feature Selection (CFS) atau Information Gain, untuk mengekstraksi atribut klinis yang memiliki korelasi tertinggi terhadap diagnosis. Langkah ini bertujuan untuk meningkatkan efisiensi komputasi model dan mempertajam akurasi hasil klasifikasi melalui penggunaan variabel data yang lebih relevan.
2. Disarankan untuk melakukan transformasi model klasifikasi ini ke dalam bentuk aplikasi yang memiliki antarmuka pengguna (User Interface) berbasis web atau mobile. Integrasi sistem klasifikasi ini ke dalam infrastruktur pelayanan kesehatan di Rumah Sakit Dirgahayu Samarinda akan memberikan manfaat praktis bagi tenaga medis dalam melakukan proses skrining awal risiko diabetes pasien secara lebih cepat dan terkomputerisasi.

5. REFERENCES

Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. Journal of Information Technology Ampera, 1(3), 133–143. https://doi.org/10.51519/journalita.volume1.isssue3.year2020.page133-143

Arrayyan, A. Z., Setiawan, H., & Putra, K. T. (2024). Naïve Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations. Semesta Teknika, 27(1), 28–36. https://doi.org/10.18196/st.v27i1.21008

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

Danny, M., & Muhidin, A. (2023). Analisis Prediksi Resiko Diabetes Tahap Awal Menggunakan Algoritma Naïve Bayes. Jurnal Teknologi Informasi Dan Komputer MH. Thamrin, 9(2), 1443–1459.

Dwi, R., Prakoso, Y., & Sari, D. R. (2022). Perancangan Sistem Klasifikasi Penyakit Diabetes. PJSE : Perwira Journal Of Science & Engineering, 02(01), 24–31.

Fansyuri, M., Ikhlas, M., Zai, S., & Mustakim, R. T. (2025). Penerapan Algoritma Naïve Bayes Untuk Mendiagnosa Penyakit Diabetes. 2(11), 2075–2078.

Karo, I. M. K., & Hendriyana. (2022). Klasifikasi Penderita Diabetes Menggunakan Algoritma Machine Learning Dan Z-Score. Jurnal Teknologi Terpadu, 8(2), 94–99.

Khasanah, L. U., Nasution, Y. N., & Amijaya, F. D. T. (2022). Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. Basis Jurnal Ilmiah Matematika, 1(1), 41–50.

Lemantara, J., & Lusiani, T. (2024). Analisis Prediksi Penyakit Diabetes Pada Wanita Menggunakan Metode Naïve Bayes Dan K-Nearest Neighbor. JITET (Jurnal Informatika Dan Teknik Elektro Terapan), 12(3).

Maulana, I., & Ernawati, S. (2025). Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan

Metode Ensemble Softvoting Dengan SMOTE-ENN dan Optimasi Bayesian. Jurnal Sains Dan Manajemen, 13(1), 71–86.

Muhammad Rafli, Z. K., & Ariatmanto, D. (2025). Analisis Perbandingan Algoritma Klasifikasi Untuk Identifikasi Diabetes Dengan Menggunakan Metode Random Forest Dan Naïve Bayes. 11–20.

Muthia, T., Nurrahma, N., & Julian, N. (2025). Seminar Nasional Ilmu Teknik dan Aplikasi Industri Penerapan teknik oversampling pada dataset indikator kesehatan diabetes menggunakan SMOTE untuk klasifikasi penderita diabetes. 8.

Nurdiana, N., & Algifari, A. (2020). Studi Komparasi Algoritma Id3 Dan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. INFOTECH Journal, 6 (2), 18–23. https://ejournal.unma.ac.id/index.php/infotech/article/view/816

Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 1–12. https://doi.org/10.1016/j.cmpb.2022.106773

Rahayu, C. A. (2023). Prediksi Penderita Diabetes Menggunakan Metode Naïve Bayes. Jurnal Informatika Dan Teknik Elektro Terapan, 11(3), 261–266. https://doi.org/10.23960/jitet.v11i3.3055

Saputro, K. E., Fadli, S., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4 . 5 , SVM , dan Naïve Bayes K-means-SMOTE for handling class imbalance in the classification of diabetes. 8(April), 89–93. https://doi.org/10.14710/jtsiskom.8.2.2020.89-93

Sulistiyowati, N., & Jajuli, M. (2020). Integrasi Naïve Bayes Dengan Teknik Sampling SMOTE Untuk Menangani Data Tidak Seimbang. Nuansa Informatika, 14(1), 34–37. https://doi.org/10.25134/nuansa.v14i1.2411

Suryanegara, G. A. B., Adiwijaya, & Purbolaksono, M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk. JURNAL RESTI, 1(10), 114–122.

Syahrullah, & Nurwijayanti, K. (2023). Klasifikasi Diagnosa Penyakit Diabetes Dengan Metode Naïve Diabetes Classification Using Web-Based Naïve Bayes Method. Jurnal Kecerdasan Buatan Dan Teknologi Informasi, 2(3), 115–121.

Widodo, Y. B., Anggraeini, S. A., Sutabri, T., & Sakaria, M. A. (2025). Perancangan Sistem Pakar Prediksi Diagnosis Penyakit Diabetes Menggunakan Algoritma Naïve Bayes Berbasis Web. Jikom: Jurnal Informatika Dan Komputer, 15(1), 161–171. https://doi.org/10.55794/jikom.v15i1.280