JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 112 – 122
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis dan Optimalisasi Performa Algoritma Gaussian Naive Bayes pada Prediksi *Metabolic Syndrome* Menggunakan SMOTE

Nadiyah Jihan Fauziyah [(1)*], Fadilla Rahmania [(2)], Muhammad Daniyal [(3)], Nur Fitriyah Ayu Tunjung Sari [(4)]
Teknik Informatika, Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim, Malang
e-mail : {nadiyahjihanf,fdrahmania1,asus.daniyal}@gmail.com, nur.fitriyah@ti.uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 28 Januari 2024, direvisi 30 April 2024, diterima 2 Mei 2024, dan dipublikasikan 25 Mei 2024.

Abstract

Metabolic syndrome is a complex global health problem, with symptoms such as abdominal obesity, insulin resistance, high blood pressure, high blood sugar, and abnormal blood lipids. With this global challenge, several studies have attempted to predict these diseases using machine learning methods. However, often, predictions about a disease result in data imbalance where minority classes are underrepresented. To balance the class proportions, the Synthetic Minority Over-sampling Technique (SMOTE) method replicates the minority class samples. In this research, the technique applied to predict is the Gaussian Naive Bayes (GNB) algorithm. The results show an increase in prediction accuracy by 0.2 from 0.81 to 0.83. This study confirms the critical role of the SMOTE oversampling method in machine learning using the Gaussian Naive Bayes (GNB) algorithm in Metabolic Syndrome prediction and its positive impact on diagnostic efficiency and public health.

Keywords: Metabolic Syndrome, Machine Learning,* Gaussian Naive Bayes*, Synthetic Minority Over-sampling Technique (SMOTE), Prediction

Abstrak

Sindrom Metabolik merupakan masalah kesehatan global yang kompleks, dengan gejala seperti obesitas abdominal, resistensi insulin, tekanan darah tinggi, gula darah tinggi, dan lipid darah abnormal. Menghadapi tantangan global ini, beberapa penelitian telah berusaha memprediksi penyakit ini dengan menggunakan metode pembelajaran mesin. Namun, seringkali prediksi tentang sebuah penyakit menghasilkan ketidakseimbangan data di mana kelas minoritas kurang terwakili. Dalam menyeimbangkan proporsi kelas, metode Synthetic Minority Over-sampling Technique (SMOTE) digunakan dengan mereplikasi sampel kelas minoritas. Dalam penelitian ini, metode yang diterapkan untuk memprediksi adalah algoritma Gaussian Naive Bayes (GNB). Hasilnya menunjukkan peningkatan akurasi prediksi sebesar 0,2 dari yang awalnya 0,81 menjadi 0,83. Penelitian ini menegaskan peran penting metode oversampling SMOTE pada pembelajaran mesin menggunakan algoritma Gaussian Naive Bayes (GNB) dalam prediksi Sindrom Metabolik dan serta dampak positifnya terhadap efisiensi diagnostik dan kesehatan masyarakat.

Kata Kunci: Sindrom Metabolik, Pembelajaran Mesin, Gaussian Naive Bayes, Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE), Prediksi

1. PENDAHULUAN

Penyakit tidak menular, terutama Sindrom Metabolik telah menjadi penyebab kematian global yang signifikan (World Health Organization, 2020). Sindrom Metabolik atau *Metabolic Syndrome* (MetS) adalah kondisi medis yang didiagnosis ketika seseorang memiliki sejumlah faktor risiko yang meningkat untuk penyakit jantung, diabetes tipe 2, dan penyakit pembuluh darah lainnya. Sindrom Metabolik dikenal sebagai himpunan gejala yang mencakup obesitas abdominal, resistensi insulin, tekanan darah tinggi, kadar gula darah tinggi, dan kadar lipid darah abnormal (Huang, 2009). Sindrom ini juga dikenal sebagai sindrom X, resistensi insulin, dll. Dalam literatur, sebenarnya ini bukan penyakit tunggal namun merupakan kumpulan faktor risiko penyakit kardiovaskular dan didefinisikan sedikit berbeda oleh berbagai organisasi (Saklayen, 2018).

Angka kejadian sindrom metabolik seringkali sejajar dengan kejadian obesitas dan kejadian diabetes tipe 2 yang merupakan salah satu akibat dari sindrom metabolik (Palaniappan et al., 2011). Menurut survei obesitas global di 195 negara yang dilakukan pada tahun 2015, 604 juta orang dewasa dan 108 juta anak-anak mengalami obesitas. Sejak tahun 1980, prevalensi obesitas meningkat dua kali lipat di 73 negara dan meningkat di sebagian besar negara lainnya. Kekhawatiran yang lebih besar adalah bahwa tingkat peningkatan obesitas pada masa kanak-kanak bahkan lebih tinggi (The GBD 2015 Obesity Collaborators, 2017). Di Indonesia sendiri, kondisi ini memiliki dampak signifikan pada kesehatan masyarakat, dengan sekitar 21,66% populasi yang didiagnosis menderita Sindrom Metabolik (Herningtyas & Ng, 2019).

Sindrom metabolik ditandai dengan masalah yang berhubungan dengan obesitas, menunjukkan adanya hubungan antara obesitas dan sindrom metabolik (Han & Lean, 2016). Faktor risiko sindrom metabolik antara lain peningkatan lingkar pinggang atau lemak perut, tingginya trigliserida plasma, peningkatan tekanan darah, gula darah tinggi, dan rendahnya *high-density lipoprotein* (HDL) (Rochlani et al., 2017). Jika seorang pasien memiliki tiga dari lima faktor risiko utama, maka pasien tersebut dikatakan mengalamai sindrom metabolik (Dobrowolski et al., 2022). Beberapa penelitian menunjukkan bahwa risiko sindrom metabolik dapat dibalik secara signifikan dengan mengurangi berat badan dan memfokuskan intervensi pada perubahan pola makan seperti pembatasan waktu makan, pola makan khusus seperti pola makan Mediterania, termasuk meningkatkan latihan fisik, mengubah pola tidur, atau bahkan mengurangi stress yang mengakibatkan sindrom metabolik (Wilkinson et al., 2020).

Beberapa pernyataan mengenai Sindrom Metabolik sebelumnya mendorong kebutuhan akan deteksi dini dan pengelolaan yang efektif. Penyakit kardiovaskular, diabetes, dan komplikasi kesehatan lainnya dapat dihindari atau dikelola lebih baik dengan adanya pendekatan preventif (Han & Lean, 2016). Keterbatasan data dan tantangan dalam prediksi Sindrom Metabolik menuntut eksplorasi terhadap algoritma-algoritma baru yang dapat memberikan kontribusi signifikan. Dari beberapa kasus prediksi penyakit, terdapat beberapa metode yang bisa digunakan, di antaranya *Logitic Regression*, *Gradient Boosting Machine* (GBM)*, K-Nearest Neighbors* (KNN), *Decision Trees* dan *Random Forests*, serta *Gaussian Naïve Bayes* (Zhou et al., 2022).

Dalam penelitian Hu et al. (2022) mereka melakukan prediksi model terhadap 2.714 (30,3%) peserta yang didiagnosis menderita sindrom metabolik. Evaluasi kinerja model menggunakan *Light Gradient Boosting Machine* (LGBM) menunjukkan hasil yang mengesankan. Model pertama memiliki nilai area *under the curve* (AUC) sebesar 0,993, sementara model kedua menunjukkan nilai AUC sebesar 0,885. Meskipun demikian, Model 3 memiliki nilai AUC sebesar 0,859, yang mendekati nilai AUC dari model kedua. Selain itu, nilai AUC untuk model *Logistic Regression* (LR) 1 dan 2 dalam skenario di rumah sakit, serta Model 3 di rumah masing-masing, adalah 0,938, 0,839, dan 0,820.

Penelitian Tavares et al. (2022) juga membahas tentang prediksi sindrom metabolik menggunakan model *machine learning* untuk memprediksi seperti*, Logistic Regression*, *Linear Discriminant Analysis*, *K-Nearest Neighbors (KNN)*, *Decision Trees*, *Light Gradient Boosting Machine (LGBM)*, dan *Extreme Gradient Boosting*. Semua model menunjukkan kalibrasi yang memadai dan diskriminasi yang baik, namun LGBM menunjukkan kinerja yang lebih baik (Sensitivitas = 87,8%, Spesifisitas = 70,2%, AUC-ROC = 0,86). Analisis inferensi kausal menunjukkan bahwa peningkatan tingkat aktivitas fisik dan pengurangan BMI setidaknya sebesar 2% memiliki efek pada pengurangan probabilitas prediksi sindrom metabolik sebesar 3,8% (95% CI = -4,8%; -2,7%).

Pada penelitian ini akan mengolah data yang bersumber di kaggle.com dan mengimplementasikan algoritma Gaussian Naive Bayes yang merupakan algoritma pembelajaran mudah yang memanfaatkan aturan *Bayes* dengan premis atau asumsi tinggi yang karakteristiknya bergantung pada kemandirian yang diberikan oleh kelas (Anand et al., 2022).

Pemilihan algoritma ini didasarkan pada kebutuhan untuk mendapatkan model yang optimal dalam menangani masalah kompleks yang terkait dengan prediksi sindrom metabolik.

Penelitian Venkata & Pandya (2022) menggunakan algoritma Gaussian Naive Bayes dalam kasus prediksi yang membandingkan beberapa model pembelajaran mesin. Dari hasil penelitian yang menerapkan algoritma *Gaussian Naïve Bayes* menghasilkan akurasi sebesar 99,66% dan standar deviasi sebesar 0,86% ketika diuji dengan menggunakan *k-means cross validation*. Sehingga pengujian secara probabilitas juga cocok untuk prediksi sindrom metabolik.

Selain itu juga terdapat penelitian Libnao et al. (2023) membuat sistem prediksi dan klasifikasi insiden lalu lintas, yang memanfaatkan algoritma *Naïve Bayes*, telah mendapat persetujuan dari Otoritas Pembangunan Metropolitan Manila karena tingkat akurasinya yang tinggi. Penelitian tersebut berhasil menunjukkan kemampuannya dalam memprediksi dan mengklasifikasikan insiden lalu lintas di Metro Manila dengan tingkat akurasi yang mengesankan sebesar 70,03%, menggunakan kumpulan data sebanyak 8.891 catatan.

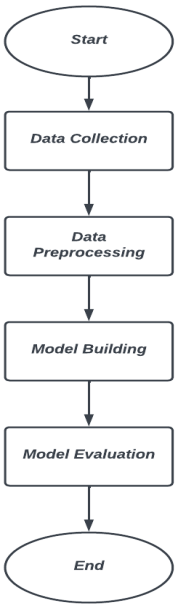
Dalam masalah di atas terdapat akurasi yang lebih rendah daripada penelitian sebelumnya. Hal ini dapat dilakukan *oversampling* pada data yang digunakan pada *machine learning*. Kombinasi metode yang melakukan *over-sampling* pada kelas minoritas (abnormal) dan *under-sampling* pada kelas mayoritas (normal) dapat mencapai kinerja pengklasifikasi yang lebih baik (dalam ruang ROC) daripada hanya melakukan *under-sampling* pada kelas mayoritas. Selain itu, metode *over-sampling* kelas minoritas dan *under-sampling* kelas mayoritas dapat mencapai kinerja pengklasifikasi yang lebih baik (dalam ruang ROC) daripada memvariasikan rasio kerugian di Ripper atau kelas *prior* di Naive Bayes (Chawla et al., 2002).

Penelitian sebelumnya menunjukkan bahwa beberapa teknik *machine learning*, termasuk algoritma Gaussian Naive Bayes, memiliki potensi untuk memprediksi sindrom metabolik dengan tingkat akurasi yang memuaskan. Namun, penelitian sebelumnya belum sepenuhnya mengeksplorasi metode yang efektif untuk mengatasi ketidakseimbangan kelas dalam data sindrom metabolik dan bagaimana hal ini dapat memengaruhi kinerja algoritma. Oleh karena itu, penelitian ini merespon kebutuhan untuk mengisi kesenjangan ini dengan mengoptimalkan algoritma Gaussian Naive Bayes secara khusus untuk prediksi sindrom metabolik, dengan tujuan meningkatkan akurasi prediksi dan relevansi klinisnya dalam konteks penanganan penyakit metabolik.

Penelitian ini bertujuan utama untuk membentuk, menganalisis, dan mengoptimalkan performa algoritma Gaussian Naive Bayes dalam prediksi sindrom metabolik. Dengan menggali potensi algoritma ini, penelitian ini berupaya memberikan kontribusi penting dalam pengembangan model deteksi dini sindrom metabolik. Dengan adanya model yang handal, diharapkan dapat meningkatkan efisiensi diagnosa, mengurangi risiko penyakit, dan secara positif memengaruhi kesehatan masyarakat secara keseluruhan.

2. METODE PENELITIAN

Pada tahap penelitian ini diberikan penjelasan sistematis mengenai urutan proses yang dilakukan dalam penelitian. Tahapan yang diuraikan dalam rangkaian ini dapat dipahami mulai dari analisis kebutuhan hingga hasil penelitian. Penelitian ini melibatkan beberapa tahapan antara lain analisis kebutuhan data, pengumpulan data, *preprocessing* data, pembangunan model menggunakan algoritma *Gaussian Naive Byes* menggunakan bahasa pemrograman Python pada platform Google Colab, evaluasi model menggunakan SMOTE, dan visualisasi hasil. Berdasarkan urutan tersebut maka tahapan penelitian akan digambarkan pada Gambar 1.



Gambar 1 Desain Sistem

2.1 *Data Collection*

Tabel 1 Data Sindrom Metabolik

	seqn	Age	Sex	Marital	Income	Race	…	BC	HDL	Tc	MS
0	62161	22	Male	Single	8200.0	White	…	92	41	84	0
1	62164	44	Female	Married	4500.0	White	…	82	28	56	0
2	62169	21	Male	Single	800.0	Asian	…	107	43	78	0
3	62199	57	Male	NaN	9000.0	White	…	100	35	98	1
4	62218	38	Female	Single	8200.0	Black	…	102	36	162	1

Tabel 2 Deskripsi Kolom

Kolom	Deskripsi
seqn	Nomor identifikasi berurutan.
Age	Usia individu.
Sex	Jenis kelamin individu (misalnya, Pria, Wanita).
Marital	Status perkawinan individu.
Income	Tingkat pendapatan atau informasi terkait pendapatan.
Race	Latar belakang etnis atau ras individu.
WaistCirc	Pengukuran lingkar pinggang.
BMI	Indeks Massa Tubuh, ukuran komposisi tubuh.
Albumunuria	Pengukuran terkait albumin dalam urin.
UrAlbCr	Rasio albumin terhadap kreatinin urin.
UricAcid	Kadar asam urat dalam darah.
BC	Kadar glukosa darah, indikator risiko diabetes.
HDL	Kadar kolesterol High-Density Lipoprotein (kolesterol "baik").
Tc	Kadar trigliserida dalam darah.
MS	Variabel biner menunjukkan ada (1) atau tidak adanya (0) sindrom metabolik.

Data yang digunakan diperoleh dari situs Kaggle (*kaggle.com*), sebuah *platform* yang menyediakan *dataset* untuk berbagai proyek *data science*. Data yang dimiliki oleh Albert Antony ini berisi informasi tentang individu dengan sindrom metabolik, suatu kondisi medis kompleks

yang terkait dengan sekelompok faktor risiko penyakit kardiovaskular dan diabetes tipe 2. Data tersebut meliputi pengukuran demografi, klinis, dan laboratorium, serta ada tidaknya sindrom metabolik. Data ini memiliki panjang 2402 data dengan 15 kolom atribut. Contoh data sindrom metabolik dapat dilihat pada Tabel 1. Adapun data pada Tabel 1 terdapat 15 kolom atribut yang dapat dideskripsikan dalam Tabel 2.

2.2 *Data Preprocessing*

Dalam tahap ini dilakukan beberapa pembersihan dan penyesuaian terhadap beberapa komponen di dalamnya seperti adanya nilai *null*, *encode* label dll. Tahap ini meliputi *Null-Handling* yang melibatkan penanganan nilai-nilai yang hilang (*null*) dalam *dataset*. *Null-handling* bisa mencakup penghapusan baris atau kolom yang mengandung nilai *null*, atau penggantian nilai *null* dengan nilai yang sesuai, seperti nilai rata-rata atau median. Pada data yang dimiliki terdapat kolom yang bersifat *Null*. Dalam proses ini dilakukan penghapusan baris dalam kolom yang terdapat nilai kosong.

Tahap selanjutnya yaitu *Label-Encoding* yang mana jika terdapat variabel kategori yang bersifat nominal atau ordinal, perlu dilakukan *label encoding* yang dapat dilihat pada Gambar 2. *Label encoding* mengubah nilai-nilai kategori menjadi angka-angka agar dapat diproses oleh algoritma *machine learning*. Tahap ini menggunakan LabelEncoder dari scikit-learn, kolom kategorikal seperti 'Marital', 'Sex', dan 'Race' diubah menjadi nilai numerik. Setelah itu, dilakukan pembuatan salinan *data frame* dengan menghapus kolom target 'MetabolicSyndrome'.

	seqn	Age	Sex	Marital	Income	Race	WaistCirc	BMI	Albuminuria	UrAlbCr	UricAcid	BloodGlucose	HDL	Triglycerides	MetabolicSyndrome
0	62161	22	Male	Single	8200.0	White	81.0	23.3	0	3.88	4.9	92	41	84	0
1	62164	44	Female	Married	4500.0	White	80.1	23.2	0	8.55	4.5	82	28	56	0
2	62169	21	Male	Single	800.0	Asian	69.6	20.1	0	5.07	5.4	107	43	78	0
3	62172	43	Female	Single	2000.0	Black	120.4	33.3	0	5.22	5.0	104	73	141	0
4	62177	51	Male	Married	NaN	Asian	81.1	20.1	0	8.13	5.0	95	43	126	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2396	71901	48	Female	Married	1000.0	Other	NaN	59.7	0	22.11	5.8	152	57	107	0
2397	71904	30	Female	Single	2000.0	Asian	NaN	18.0	0	2.90	7.9	91	90	91	0
2398	71909	28	Male	Single	800.0	MexAmerican	100.8	29.4	0	2.78	6.2	99	47	84	0
2399	71911	27	Male	Married	8200.0	MexAmerican	106.6	31.3	0	4.15	6.2	100	41	124	1
2400	71915	60	Male	Single	6200.0	White	106.6	27.5	0	12.82	5.2	91	36	226	1

2401 rows × 15 columns

Gambar 2 Data Sebelum *Preprocessing*

Sebelum *preprocessing*, terdapat 2401 *record* data yang terlihat pada Gambar 2. Setelah proses tersebut, jumlah data berkurang menjadi 2009 record. *Preprocessing* penting untuk membersihkan dan mempersiapkan data sebelum analisis lebih lanjut.

	seqn	Age	Sex	Marital	Income	Race	WaistCirc	BMI	Albuminuria	UrAlbCr	UricAcid	BloodGlucose	HDL	Triglycerides	MetabolicSyndrome
0	62161	22	1	3	8200.0	5	81.0	23.3	0	3.88	4.9	92	41	84	0
1	62164	44	0	1	4500.0	5	80.1	23.2	0	8.55	4.5	82	28	56	0
2	62169	21	1	3	800.0	0	69.6	20.1	0	5.07	5.4	107	43	78	0
3	62172	43	0	3	2000.0	1	120.4	33.3	0	5.22	5.0	104	73	141	0
5	62178	80	1	4	300.0	5	112.5	28.5	0	9.79	4.8	105	47	100	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2394	71895	31	1	1	2500.0	0	74.0	20.6	0	2.00	6.7	95	64	81	0
2395	71898	65	0	1	5400.0	3	98.5	29.4	0	5.51	6.7	114	49	165	1
2398	71909	28	1	3	800.0	3	100.8	29.4	0	2.78	6.2	99	47	84	0
2399	71911	27	1	1	8200.0	3	106.6	31.3	0	4.15	6.2	100	41	124	1
2400	71915	60	1	3	6200.0	5	106.6	27.5	0	12.82	5.2	91	36	226	1

2009 rows × 15 columns

Gambar 3 Data Setelah *Preprocessing*

2.3 *Model Building*

Setelah data sudah bersih dan siap, selanjutnya adalah tahap pembuatan model dengan mengimplementasikan algoritma Gaussian Naive Bayes pada *dataset* yang telah diproses sebelumnya. Terdapat beberapa hal yang dilakukan dalam tahap ini, yang pertama yaitu pembagian data menjadi dua bagian yaitu menjadi data latih (*training data*) dan data uji (*testing data*). Data latih digunakan untuk melatih model, sementara data uji digunakan untuk menguji performa model pada data yang belum pernah dilihat sebelumnya.

Selanjutnya yaitu tahap impelementasi algoritma Gaussian Naive Bayes. Algoritma ini sama seperti teorema bayes di mana persamaannya disajikan pada Pers. (1). Di mana P(A) merupakan nilai probabilitas A, P(B) nilai probabilitas B, P(A|B) nilai probabilitas A menyebabkan B (probabilitas posterior), dan P(B|A) nilai probabilitas B menyebabkan A.

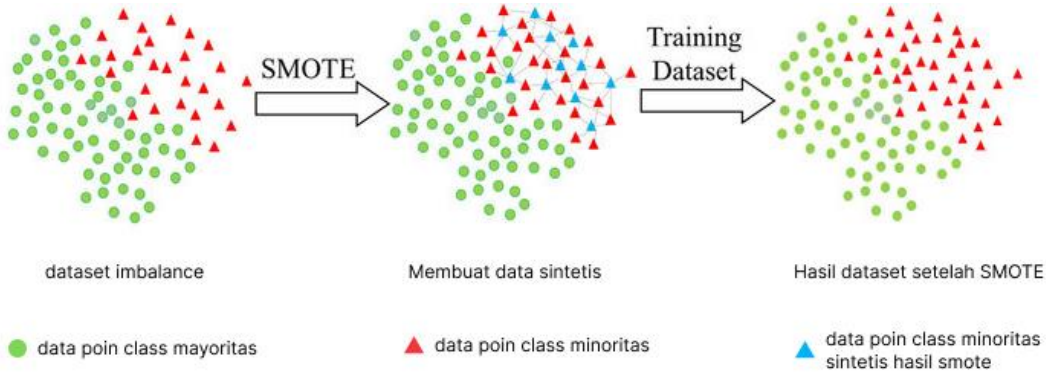
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

Perbedaannya terletak pada perhitungan probabilitas tiap fiturnya menggunakan persamaan distribusi normal seperti yang ditampilkan pada Pers. (2). Di mana P(x(A) | y(B)) merupakan nilai probabilitas fitur x(A) terhadap target y(B), *x* yaitu nilai fitur yang di prediksi, σ nilai Standar deviasi, dan μ merupakan nilai rata-rata.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2}$$

2.4 *Model Evaluation*

Setelah model berhasil dilatih dan diuji, tahapan selanjutnya melibatkan analisis performa model untuk mengukur sejauh mana keefektifan prediksi. Dalam analisis ini, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-*score* dievaluasi sesuai dengan jenis masalah klasifikasi yang dihadapi. Pentingnya memahami metrik-metrik ini adalah untuk mendapatkan wawasan yang komprehensif tentang seberapa baik model dapat mengklasifikasikan data. Selanjutnya, untuk mengatasi ketidakseimbangan kelas dalam *dataset*, dilakukan optimalisasi dengan menerapkan teknik *class balancing* menggunakan SMOTE (Synthetic Minority Over-sampling Technique).



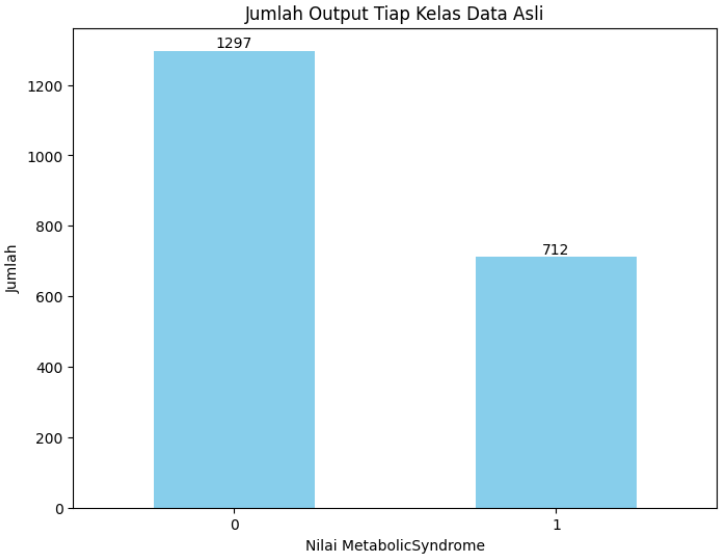
Gambar 4 Visualisasi Cara Kerja SMOTE

Jika terdapat ketidakseimbangan dalam distribusi kelas pada data, teknik *class balancing* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dapat digunakan. Cara kerja SMOTE ialah menciptakan sampel-sampel sintetis dari kelas minoritas untuk menyamakan jumlah sampel antara kelas mayoritas dan minoritas seperti yang ditampilkan dalam Gambar 4. Hal ini

membantu mencegah model menjadi bias terhadap kelas mayoritas dan meningkatkan kemampuan model dalam mengenali kelas minoritas.

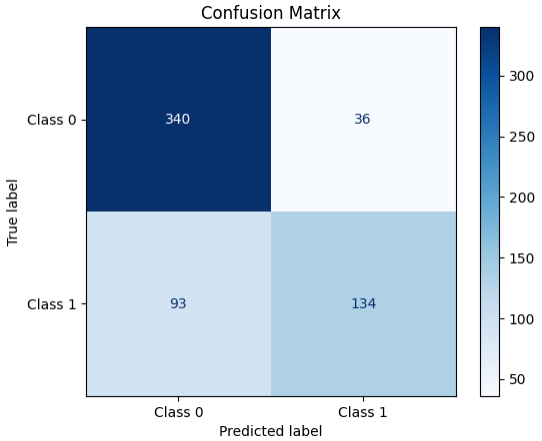
3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pengujian sistem dilakukan dengan membagi *dataset* menjadi dua bagian utama yaitu 70% data digunakan untuk data *training*, sementara 30% sisanya digunakan untuk menguji performa model atau data *testing*. Pembagian ini bertujuan untuk menghasilkan model klasifikasi yang dapat memahami dan mengklasifikasikan label penyakit *Metabolic Syndrome* menjadi "Ya" atau "Tidak". Pembagian kelas pada data dapat dilihat pada Gambar 5. Terlihat dari Gambar 5, bahwa terjadi ketidakseimbangan kelas pada nilai *Metabolic Syndrome*, sehingga dibutuhkan adanya sebuah prediksi untuk membuktikan apakah data tersebut benar valid atau tidak.



Gambar 5 Pembagian Kelas pada Data *Metabolic Syndrome*

3.1 Hasil Pengujian Menggunakan Gaussian Naïve Bayes



Gambar 6 *Confusion Matrix* Menggunakan Gaussian Naïve Bayes

Pada percobaan pertama yang menggunakan algoritma Gaussian Naive Bayes dapat dilihat pada Gambar 6 yang menghasilkan nilai akurasi, presisi, *recall*, dan F1-*Score* pada Tabel 3.

Setelah dilakukan evaluasi dari model, didapat perbedaan *recall* yang cukup signifikan antara kelas “0” dan kelas “1”. Perbedaan *recall* yang cukup signifikan ini mengindikasikan bahwa model sangat baik dalam menentukan kelas “0” dan kurang baik dalam menentukan kelas “1”. Hal ini terjadi dikarenakan ketidakseimbangan antara data kelas “0” dan “1” pada data pelatihan, sehingga perlu dilakukan penyeimbangan jumlah kelas pada data pelatihan.

Tabel 3 Nilai dari *Confusion Matrix* GNB

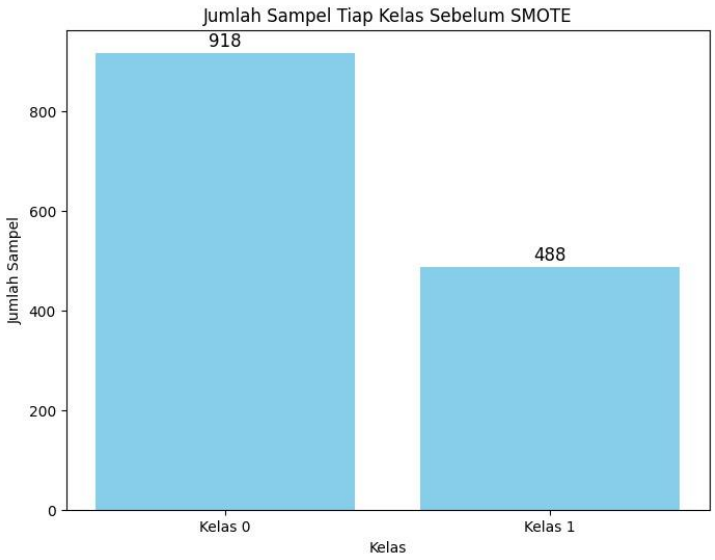
	Precision	***Recall***	***F1-Score***
0	0.81	0.93	0.86
1	0.82	0.59	0.68
Accuracy			**0.81**

3.2 Hasil Pengujian Menambahkan Teknik SMOTE

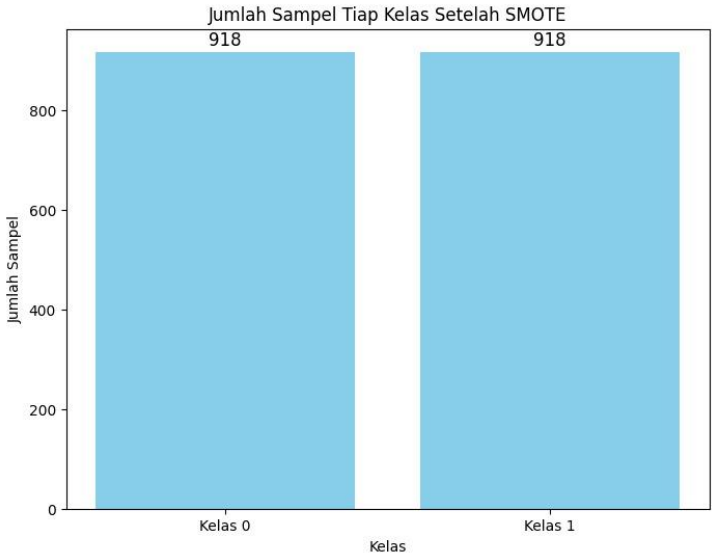
Salah satu cara untuk mengatasi data kelas yang tidak seimbang adalah dengan dilakukan *oversampling* pada data. Teknik *oversampling* yang umum dipakai adalah SMOTE seperti pada Gambar 4. Adapun proses SMOTE yang digunakan menggunakan *library* ’imblearn.over_sampling import SMOTE’, lalu diinisialisasi dengan mengatur random_state ke nilai tertentu untuk memastikan reproduksibilitas dengan menambahkan *code* ’sm = SMOTE(random_state=2)’. Setelah itu *oversampling* diterapkan dengan penggunaan metode ’*fit_resample’* menjadi fungsi di bawah ini:

```
x_resampled, y_resampled = sm.fit_resample(x_train, y_train.ravel())
```

Proses ini akan menciptakan contoh sintetis baru dari kelas minoritas (kelas yang jumlahnya lebih sedikit) sehingga jumlah sampel dalam kedua kelas menjadi seimbang. Proses ini membuat *dataset* menjadi lebih seimbang dan membantu model untuk mempelajari pola dari kedua kelas dengan lebih baik, yang diharapkan akan meningkatkan kinerja model dalam melakukan klasifikasi pada kelas minoritas. Adapun perbedaan pembagian kelas dari sebelum dan setelah menggunakan SMOTE dijelaskan pada gambar 7 dan 8. Setelah dilakukan penyeimbangan kelas dengan *oversampling* menggunakan SMOTE dapat dilihat pada Gambar 9 yang menghasilkan nilai akurasi, presisi, *recall*, dan F1-*Score* pada Tabel 4.



Gambar 7 Pembagian Kelas Sebelum SMOTE



Gambar 8 Pembagian Kelas Setelah SMOTE

Gambar 9 *Confusion Matrix* Menggunakan Performa SMOTE

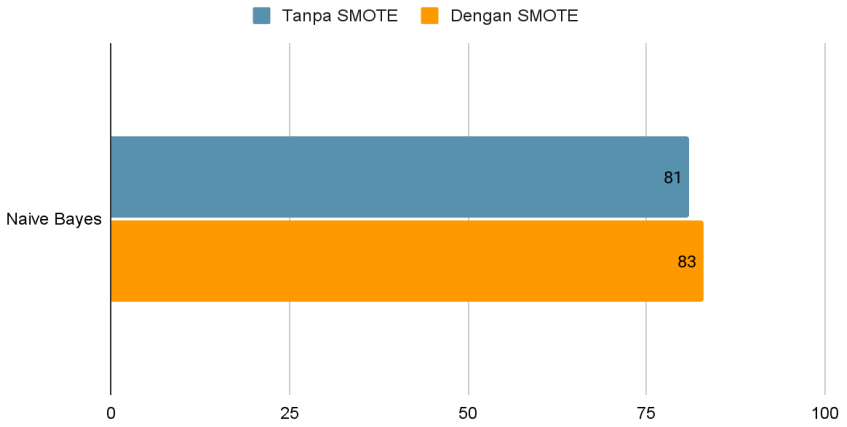
Tabel 4 Nilai dari *Confusion Matrix* Performa SMOTE

	Precision	***Recall***	***F1-Score***
0	0.86	0.88	0.87
1	0.78	0.73	0.75
Accuracy			**0.83**

Berdasarkan pelatihan model setelah *oversampling* dengan SMOTE, akurasi dari model mengalami kenaikan sebanyak 2% menjadi 83%. Terdapat beberapa perubahan dalam *classification report* kedua. Kenaikan terjadi pada *recall* dan F1-*score* serta terjadi sedikit penurunan pada *precision* terutama pada kelas “1”. Perbandingan akurasi ini juga dapat dilihat pada Gambar 10.

Kenaikan akurasi sebesar 2% dari 81% menjadi 83% pada model kedua yang menggunakan Gaussian Naive Bayes (GNB) setelah penerapan SMOTE mengindikasikan peningkatan yang signifikan dalam kemampuan model untuk mengklasifikasikan kasus sindrom metabolik. SMOTE memberikan kontribusi penting dengan meningkatkan *recall* pada kelas minoritas, yang dapat menjadi kritis dalam aplikasi klinis di mana deteksi yang tepat dari kasus positif sangat penting.

Selain itu, peningkatan F1-*Score* menunjukkan peningkatan keseluruhan dalam kemampuan model untuk melakukan klasifikasi yang akurat. Sampai saat ini belum ada penelitian yang menggunakan teknik SMOTE pada penerapan metode Gaussian Naive Bayes, terlebih lagi pada prediksi Metabolic Syndrome.



Gambar 10 Perbandingan Akurasi

4. KESIMPULAN

Selama pengujian model pertama, diperoleh akurasi sebesar 81%, mengindikasikan kemampuan model dalam mengklasifikasikan sebagian besar kasus secara benar. *Precision* sebesar 88.5% menunjukkan tingkat keakuratan model ketika menyatakan seseorang memiliki sindrom metabolik. *Recall* sebesar 85% menggambarkan kemampuan model untuk menangkap sebagian besar kasus sindrom metabolik yang sebenarnya. F1-*Score* mencapai 86.7%, mencerminkan keseimbangan yang baik antara kemampuan model untuk mengidentifikasi kasus positif dan menghindari *false positive*.

Pada pelatihan model kedua, terjadi peningkatan akurasi menjadi 83%, dengan *recall* pada kelas "1" meningkat dari 0.59 menjadi 0.73. Hal ini menandakan peningkatan signifikan dalam kemampuan model untuk mendeteksi kasus sebenarnya dari kelas "1" dibandingkan dengan model sebelumnya. Meskipun terdapat sedikit penurunan *recall* pada kelas "0", namun hal ini masih dapat diterima. Peningkatan 7% pada F1-*Score* pada kelas "1" dan 1% pada kelas "0" menunjukkan peningkatan performa model dalam mengidentifikasi kasus positif dan menghindari *false positive*.

Hasil ini memberikan dorongan untuk pengembangan lebih lanjut dalam mendekati deteksi Sindrom Metabolik dengan menggunakan metode *machine learning*, dengan potensi peningkatan performa melalui penyesuaian dan peningkatan teknik seperti *oversampling* menggunakan SMOTE. Keseluruhan, penelitian ini memberikan kontribusi pada pemahaman mendalam tentang aplikasi model *machine learning* untuk prediksi Sindrom Metabolik dan memberikan landasan untuk penelitian lebih lanjut dalam pengembangan model yang lebih canggih.

DAFTAR PUSTAKA

Anand, M. V., KiranBala, B., Srividhya, S. R., C., K., Younus, M., & Rahman, M. H. (2022). Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer. *Mobile Information Systems*, *2022*, 1–7. https://doi.org/10.1155/2022/2436946

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Dobrowolski, P., Prejbisz, A., Kuryłowicz, A., Baska, A., Burchardt, P., Chlebus, K., Dzida, G., Jankowski, P., Jaroszewicz, J., Jaworski, P., Kamiński, K., Kapłon-Cieślicka, A., Klocek, M., Kukla, M., Mamcarz, A., Mastalerz-Migas, A., Narkiewicz, K., Ostrowska, L., Śliż, D., … Bogdański, P. (2022). Metabolic syndrome – a new definition and management guidelines A joint position paper by the Polish Society of Hypertension, Polish Society for the Treatment of Obesity, Polish Lipid Association, Polish Association for Study of Liver, Polish Society of Family Medicine, Polish Society of Lifestyle Medicine, Division of Prevention and Epidemiology Polish Cardiac Society, "Club 30" Polish Cardiac Society, and Division of Metabolic and Bariatric Surgery Society of Polish Surgeons. *Archives of Medical Science*, *18*(5), 1133–1156. https://doi.org/10.5114/aoms/152921

Han, T. S., & Lean, M. E. (2016). A clinical perspective of obesity, metabolic syndrome and cardiovascular disease. *JRSM Cardiovascular Disease*, *5*, 204800401663337. https://doi.org/10.1177/2048004016633371

Herningtyas, E. H., & Ng, T. S. (2019). Prevalence and distribution of metabolic syndrome and its components among provinces and ethnic groups in Indonesia. *BMC Public Health*, *19*(1), 1–12. https://doi.org/10.1186/S12889-019-6711-7/FIGURES/3

Hu, X., Li, X.-K., Wen, S., Li, X., Zeng, T.-S., Zhang, J.-Y., Wang, W., Bi, Y., Zhang, Q., Tian, S.-H., Min, J., Wang, Y., Liu, G., Huang, H., Peng, M., Zhang, J., Wu, C., Li, Y.-M., Sun, H., … Chen, L.-L. (2022). Predictive modeling the probability of suffering from metabolic syndrome using machine learning: A population-based study. *Heliyon*, *8*(12), e12343. https://doi.org/10.1016/j.heliyon.2022.e12343

Huang, P. L. (2009). A comprehensive definition for metabolic syndrome. *Disease Models & Mechanisms*, *2*(5–6), 231–237. https://doi.org/10.1242/dmm.001180

Libnao, M., Misula, M., Andres, C., Mariñas, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve bayes algorithm. *Procedia Computer Science*, *227*, 316–325. https://doi.org/10.1016/j.procs.2023.10.530

Palaniappan, L. P., Wong, E. C., Shin, J. J., Fortmann, S. P., & Lauderdale, D. S. (2011). Asian Americans have greater prevalence of metabolic syndrome despite lower body mass index. *International Journal of Obesity*, *35*(3), 393–400. https://doi.org/10.1038/ijo.2010.152

Rochlani, Y., Pothineni, N. V., Kovelamudi, S., & Mehta, J. L. (2017). Metabolic syndrome: Pathophysiology, management, and modulation by natural compounds. *Therapeutic Advances in Cardiovascular Disease*, *11*(8), 215–225. https://doi.org/10.1177/1753944717711379/ASSET/IMAGES/LARGE/10.1177_1753944717711379-FIG1.JPEG

Saklayen, M. G. (2018). The Global Epidemic of the Metabolic Syndrome. *Current Hypertension Reports*, *20*(2), 1–8. https://doi.org/10.1007/S11906-018-0812-Z/METRICS

Tavares, L. D., Manoel, A., Donato, T. H. R., Cesena, F., Minanni, C. A., Kashiwagi, N. M., da Silva, L. P., Amaro, E., & Szlejf, C. (2022). Prediction of metabolic syndrome: A machine learning approach to help primary prevention. *Diabetes Research and Clinical Practice*, *191*, 110047. https://doi.org/10.1016/j.diabres.2022.110047

The GBD 2015 Obesity Collaborators. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine*, *377*(1), 13–27. https://doi.org/10.1056/NEJMOA1614362/SUPPL_FILE/NEJMOA1614362_DISCLOSURES.PDF

Venkata, P., & Pandya, V. (2022). Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks. *Materials Today: Proceedings*, *62*(P13), 7156–7161. https://doi.org/10.1016/j.matpr.2022.03.035

Wilkinson, M. J., Manoogian, E. N. C., Zadourian, A., Lo, H., Fakhouri, S., Shoghi, A., Wang, X., Fleischer, J. G., Navlakha, S., Panda, S., & Taub, P. R. (2020). Ten-Hour Time-Restricted Eating Reduces Weight, Blood Pressure, and Atherogenic Lipids in Patients with Metabolic Syndrome. *Cell Metabolism*, *31*(1), 92-104.e5. https://doi.org/10.1016/j.cmet.2019.11.004

World Health Organization. (2020). *The top 10 causes of death*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Zhou, Y., Wu, T., Jiang, Y., Li, Y., Li, K., Quan, L., & Lyu, Q. (2022). DeepNup: Prediction of Nucleosome Positioning from DNA Sequences Using Deep Neural Network. *Genes*, *13*(11), 1983. https://doi.org/10.3390/genes13111983