

# Analisis Kinerja Random Forest, Gradient Boosting, dan LightGBM dengan SMOTENC pada Klasifikasi Tingkat Obesitas

Kartika Handayani<sup>1\*</sup>, Erni<sup>2</sup>, Fuad Nur Hasan<sup>3</sup>

<sup>1,2</sup>Fakultas Teknik dan Informatika, Program Studi Informatika Kampus Kota Pontianak, Universitas Bina Sarana Informatika, Pontianak, Indonesia

<sup>3</sup>Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Jakarta, Indonesia

Email: <sup>1\*</sup>[kartika.kth@bsi.ac.id](mailto:kartika.kth@bsi.ac.id), <sup>2</sup>[erni.erni@bsi.ac.id](mailto:erni.erni@bsi.ac.id), <sup>3</sup>[fuad.fnu@bsi.ac.id](mailto:fuad.fnu@bsi.ac.id)

(\*Email Corresponding Author: [kartika.kth@bsi.ac.id](mailto:kartika.kth@bsi.ac.id))

Received: May 25, 2026 | Revision: June 2, 2026 | Accepted: June 19, 2026

## Abstrak

Klasifikasi tingkat obesitas merupakan aspek penting dalam bidang kesehatan untuk mendukung deteksi dini dan pengambilan keputusan yang tepat. Namun, dataset obesitas umumnya memiliki distribusi kelas yang tidak seimbang, yang dapat menyebabkan bias model terhadap kelas mayoritas dan menurunkan kemampuan dalam mengenali kelas minoritas. Penelitian ini bertujuan untuk menganalisis kinerja algoritma *ensemble learning* dalam klasifikasi tingkat obesitas serta mengevaluasi pengaruh teknik penyeimbangan data menggunakan *Synthetic Minority Oversampling Technique for Nominal and Continuous* (SMOTENC). Kontribusi penelitian ini terletak pada perbandingan pengaruh SMOTENC terhadap algoritma *ensemble learning* berbasis *bagging* dan *boosting*, yaitu *Random Forest*, *Gradient Boosting*, dan *Light Gradient Boosting Machine* (LightGBM), pada kasus klasifikasi obesitas multikelas. Dataset yang digunakan merupakan *Obesity Levels Dataset* yang terdiri dari 2.111 data, 16 atribut prediktor, dan 7 kelas target tingkat obesitas. Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta didukung oleh analisis *confusion matrix* dan kurva *Receiver Operating Characteristic* (ROC). Hasil penelitian menunjukkan bahwa model LightGBM dengan SMOTENC menghasilkan performa terbaik dengan nilai *accuracy* 96,45%, *precision* 96,54%, *recall* 96,45%, dan *F1-score* 96,47%. Selain itu, penerapan SMOTENC menunjukkan pengaruh yang berbeda pada setiap algoritma, dengan peningkatan performa paling terlihat pada model LightGBM.

**Kata Kunci:** klasifikasi obesitas, ketidakseimbangan data, *ensemble learning*, SMOTENC, LightGBM

## Abstract

*Obesity level classification is an important aspect of healthcare that supports early detection and informed decision-making. However, obesity datasets commonly exhibit imbalanced class distributions, which can lead to model bias toward majority classes and reduce the ability to recognize minority classes. This study aims to analyze the performance of ensemble learning algorithms for obesity level classification and evaluate the effect of data balancing using the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTENC). The contribution of this study lies in comparing the impact of SMOTENC on bagging- and boosting-based ensemble learning algorithms, namely Random Forest, Gradient Boosting, and Light Gradient Boosting Machine (LightGBM), in a multiclass obesity classification task. The study utilized the Obesity Levels Dataset, which consists of 2,111 instances, 16 predictive attributes, and 7 obesity-level target classes. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics, supported by confusion matrix and Receiver Operating Characteristic (ROC) curve analyses. The results indicate that the LightGBM model with SMOTENC achieved the best performance, with an accuracy of 96.45%, precision of 96.54%, recall of 96.45%, and F1-score of 96.47%. Furthermore, the application of SMOTENC demonstrated varying effects across the evaluated algorithms, with the most notable performance improvement observed in the LightGBM model.*

**Keywords:** obesity classification, class imbalance, ensemble learning, SMOTENC, LightGBM

## 1. PENDAHULUAN

Obesitas merupakan salah satu permasalahan kesehatan global yang prevalensinya terus meningkat dan menjadi faktor risiko utama berbagai penyakit tidak menular, seperti diabetes melitus tipe 2, hipertensi, penyakit kardiovaskular, gangguan metabolik, serta beberapa jenis kanker. Kondisi obesitas ditandai dengan akumulasi lemak tubuh secara berlebihan yang dapat menurunkan kualitas hidup serta meningkatkan angka morbiditas dan mortalitas masyarakat [1]. Peningkatan prevalensi obesitas dipengaruhi oleh perubahan pola hidup, rendahnya aktivitas fisik, konsumsi makanan tinggi kalori, serta perkembangan gaya hidup sedentary yang semakin meningkat pada berbagai kelompok usia [2].

Berdasarkan laporan kesehatan global, jumlah individu dengan obesitas mengalami peningkatan signifikan dalam beberapa dekade terakhir dan diproyeksikan terus bertambah pada tahun-tahun mendatang [3]. Kondisi tersebut menjadi perhatian penting karena obesitas tidak hanya berdampak pada kesehatan individu, tetapi juga meningkatkan beban ekonomi sistem pelayanan kesehatan. Oleh karena itu, identifikasi dan klasifikasi tingkat obesitas secara dini menjadi langkah strategis dalam mendukung upaya pencegahan serta pengendalian risiko penyakit yang ditimbulkan [1].

Seiring berkembangnya teknologi informasi, penerapan data mining dan machine learning pada bidang kesehatan semakin banyak digunakan sebagai pendekatan untuk membantu proses prediksi dan klasifikasi penyakit secara otomatis. Machine learning memiliki kemampuan dalam mempelajari pola dari data historis sehingga dapat digunakan untuk menghasilkan prediksi berdasarkan karakteristik tertentu [4]. Pada kasus klasifikasi tingkat obesitas, pendekatan machine

learning mampu memanfaatkan berbagai atribut seperti usia, tinggi badan, berat badan, pola konsumsi makanan, aktivitas fisik, serta kebiasaan hidup individu untuk menghasilkan klasifikasi yang lebih cepat dan objektif dibandingkan pendekatan konvensional [3].

Berbagai penelitian sebelumnya telah menerapkan algoritma data mining untuk klasifikasi data kesehatan maupun prediksi obesitas, di antaranya *Light Gradient Boosting Machine*, *Logistic Regression*, *Random Forest*, dan *Stochastic Gradient Descent (SGD)* [3]. Hasil penelitian menunjukkan bahwa pendekatan *ensemble learning* cenderung menghasilkan performa klasifikasi yang lebih baik dibandingkan algoritma tunggal karena mampu meningkatkan stabilitas model dan mengurangi risiko overfitting [5]. *Ensemble learning* terdiri atas dua pendekatan utama yaitu bagging dan boosting. *Random Forest* merupakan algoritma berbasis bagging yang bekerja melalui kombinasi beberapa decision tree, sedangkan *Gradient Boosting* dan *LightGBM* termasuk metode boosting yang membangun model secara bertahap dengan memperbaiki kesalahan prediksi sebelumnya [6].

Meskipun algoritma *ensemble learning* memiliki kemampuan klasifikasi yang tinggi, salah satu tantangan utama pada dataset kesehatan adalah ketidakseimbangan distribusi kelas (*imbalanced data*). Ketidakseimbangan data menyebabkan model cenderung lebih optimal dalam mengenali kelas mayoritas dibandingkan kelas minoritas sehingga dapat menurunkan kemampuan klasifikasi, khususnya pada nilai *recall* dan *F1-score* [7]. Kondisi tersebut berpotensi menghasilkan bias prediksi meskipun model memperoleh nilai *accuracy* yang tinggi [8].

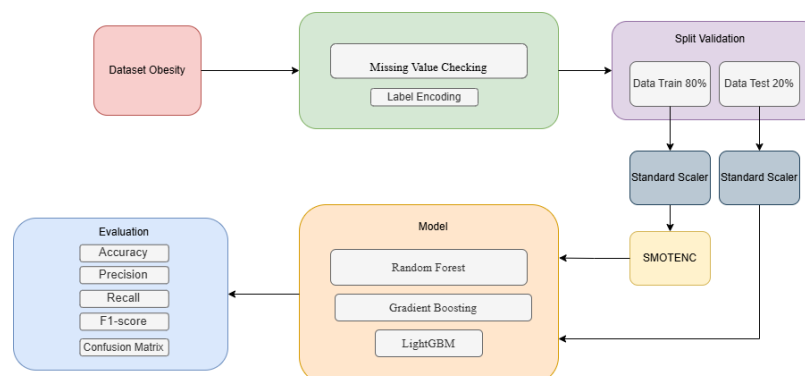
Berbagai pendekatan telah dikembangkan untuk mengatasi permasalahan *imbalanced data*, salah satunya melalui teknik *oversampling*. *Synthetic Minority Oversampling Technique (SMOTE)* merupakan metode yang umum digunakan untuk membentuk data sintesis pada kelas minoritas. Namun, metode tersebut memiliki keterbatasan dalam menangani dataset dengan atribut campuran numerik dan kategorikal [8]. Oleh karena itu, dikembangkan *Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTENC)* yang dirancang untuk menangani atribut kategorikal dan numerik secara bersamaan sehingga lebih sesuai diterapkan pada dataset kesehatan yang memiliki karakteristik heterogen [9].

Beberapa penelitian melaporkan bahwa penerapan teknik *oversampling* mampu meningkatkan kemampuan model dalam mendeteksi kelas minoritas serta memperbaiki nilai *recall* dan *F1-score* [10]. Akan tetapi, peningkatan performa tersebut tidak selalu konsisten pada seluruh algoritma klasifikasi karena dipengaruhi oleh karakteristik data dan model yang digunakan. Selain itu, penelitian mengenai klasifikasi tingkat obesitas masih didominasi oleh perbandingan performa algoritma machine learning tanpa mengkaji secara lebih mendalam pengaruh penerapan SMOTENC terhadap metode *ensemble learning*, khususnya pada pendekatan bagging dan boosting.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk melakukan evaluasi kinerja algoritma *Random Forest*, *Gradient Boosting*, dan *LightGBM* dengan penerapan SMOTENC pada klasifikasi tingkat obesitas. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menganalisis pengaruh penanganan ketidakseimbangan data terhadap performa model klasifikasi. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode klasifikasi obesitas berbasis machine learning serta memberikan pemahaman mengenai efektivitas penerapan SMOTENC pada algoritma *ensemble learning*.

## 2. METODOLOGI PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan, dimulai dari pengumpulan dataset, dilanjutkan dengan preprocessing data yang meliputi transformasi data, pembagian data latih dan data uji, standardisasi, serta penanganan ketidakseimbangan kelas menggunakan SMOTENC. Tahap berikutnya yaitu pemodelan menggunakan algoritma *Random Forest*, *Gradient Boosting*, dan *LightGBM*, kemudian dilakukan evaluasi performa model berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score* dan *confusion matrix*. Tahap akhir berupa analisis hasil klasifikasi untuk mengetahui pengaruh penerapan SMOTENC terhadap performa masing-masing algoritma dalam prediksi tingkat obesitas. Adapun alur tahapan penelitian secara keseluruhan ditunjukkan pada Gambar 1.



**Gambar 1.** Alur dan Desain Penelitian

## 2.1 Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari dataset *Obesity Levels* yang tersedia pada platform Kaggle. Dataset tersebut digunakan untuk melakukan klasifikasi tingkat obesitas berdasarkan kebiasaan makan (*eating habits*) dan kondisi fisik individu. Dataset berasal dari penelitian mengenai estimasi tingkat obesitas pada individu di negara Mexico, Peru, dan Colombia dengan rentang usia antara 14–61 tahun. Dataset terdiri dari 2.111 record dengan 17 atribut dan 1 variabel target (*NObeyesdad*) yang merepresentasikan tingkat obesitas seseorang. Sebanyak 23% data diperoleh secara langsung melalui survei berbasis web, sedangkan 77% data dihasilkan secara sintesis menggunakan metode SMOTE pada perangkat Weka untuk memperkaya distribusi data. Adapun atribut yang digunakan pada penelitian ditampilkan pada Tabel 1.

**Tabel 1.** Deskripsi Variabel Dataset

Variabel	Tipe Data	Keterangan
<i>Gender</i>	Kategorikal	Jenis kelamin
<i>Age</i>	Numerik	Umur responden
<i>Height</i>	Numerik	Tinggi badan (meter)
<i>Weight</i>	Numerik	Berat badan (kg)
<i>family_history_with_overweight</i>	Biner	Riwayat keluarga obesitas
FAVC	Biner	Konsumsi makanan tinggi kalori
FCVC	Numerik	Frekuensi konsumsi sayur
NCP	Numerik	Jumlah makan utama per hari
CAEC	Kategorikal	Kebiasaan makan di luar waktu utama
SMOKE	Biner	Kebiasaan merokok
CH2O	Numerik	Konsumsi air harian
SCC	Biner	Pemantauan kalori
FAF	Numerik	Aktivitas fisik
TUE	Numerik	Waktu penggunaan perangkat teknologi
CALC	Kategorikal	Konsumsi alkohol
MTRANS	Kategorikal	Jenis transportasi
NObeyesdad	Target	Tingkat obesitas

## 2.2 Preprocessing Data

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data dan menyesuaikan format data sebelum proses pemodelan machine learning [11]. Dataset obesitas yang digunakan terdiri dari atribut numerik dan kategorikal sehingga diperlukan beberapa tahapan transformasi data. Menurut [3], *preprocessing* berperan penting dalam meningkatkan performa model klasifikasi karena dapat mengurangi noise dan inkonsistensi data. Adapun tahapan *preprocessing* yang dilakukan pada penelitian ini yaitu:

### 1. Pemeriksaan Data (*Data Checking*)

Tahap awal dilakukan pemeriksaan terhadap *missing value* dan data duplikat. Berdasarkan hasil identifikasi, dataset tidak memiliki nilai kosong sehingga seluruh data dapat digunakan dalam proses penelitian.

### 2. Transformasi Data Kategorikal (*Label Encoding*)

Atribut kategorikal seperti Gender, CAEC, CALC, dan MTRANS diubah menjadi bentuk numerik menggunakan Label Encoding. Transformasi dilakukan karena algoritma machine learning tidak dapat memproses data berbentuk teks secara langsung [12].

### 3. Standardisasi Data (*StandardScaler*)

Normalisasi atribut numerik dilakukan menggunakan StandardScaler dengan tujuan menyamakan rentang nilai antar fitur sehingga proses pembelajaran model menjadi lebih stabil [13].

### 4. Pembagian Data Training dan Testing

Dataset dibagi menggunakan metode *train-test split* dengan proporsi 80% data training dan 20% data testing. Pembagian dilakukan menggunakan parameter *stratify* agar distribusi kelas target tetap terjaga pada kedua kelompok data [10]. Data *training* digunakan untuk membangun model klasifikasi, sedangkan data *testing* digunakan untuk mengevaluasi performa model.

### 2.3 Penanganan Ketidakseimbangan Data Menggunakan SMOTENC

Ketidakseimbangan distribusi kelas (*imbalanced data*) dapat menyebabkan model cenderung mengenali kelas mayoritas dibandingkan kelas minoritas [14]. Oleh karena itu, penelitian ini menerapkan *Synthetic Minority Oversampling Technique for Nominal and Continuous* (SMOTENC). SMOTENC merupakan pengembangan metode SMOTE yang dirancang untuk menangani data campuran numerik dan kategorikal [8]. Pada penelitian ini, atribut yang ditetapkan sebagai fitur kategorikal dalam proses SMOTENC meliputi Gender, CALC, FAVC, SCC, SMOKE, family\_history\_with\_overweight, CAEC, dan MTRANS. Penetapan atribut tersebut dilakukan agar proses pembangkitan data sintesis pada kelas minoritas dapat mempertahankan karakteristik data kategorikal sekaligus menghindari terbentuknya nilai kategori yang tidak valid.

### 2.4 Pemodelan Ensemble Learning

Penelitian ini menggunakan tiga algoritma ensemble learning yaitu Random Forest, Gradient Boosting, dan LightGBM. Ensemble learning diketahui mampu meningkatkan stabilitas model serta mengurangi kesalahan prediksi [16].

#### 1. Random Forest

Random Forest merupakan metode berbasis bagging yang membangun sejumlah *decision tree* kemudian menghasilkan prediksi akhir melalui *majority voting* [17].

#### 2. Gradient Boosting

Gradient Boosting merupakan metode boosting yang bekerja dengan memperbaiki kesalahan model sebelumnya secara bertahap sehingga menghasilkan performa prediksi lebih baik [18].

#### 3. LightGBM

LightGBM (*Light Gradient Boosting Machine*) merupakan algoritma boosting berbasis *decision tree* yang dirancang untuk meningkatkan efisiensi komputasi serta mempercepat proses pelatihan model [6].

Pengujian dilakukan menggunakan dua skenario:

1. Model tanpa SMOTENC
2. Model dengan SMOTENC

Sehingga diperoleh:

- Random Forest
- Random Forest + SMOTENC
- Gradient Boosting
- Gradient Boosting + SMOTENC
- LightGBM
- LightGBM + SMOTENC

### 2.5 Evaluasi Model

Evaluasi model dilakukan menggunakan *confusion matrix* dengan beberapa metrik pengukuran yaitu *accuracy*, *precision*, *recall*, dan *F1-score* [19].

1. Accuracy  
Accuracy menunjukkan tingkat ketepatan prediksi model terhadap seluruh data.
2. Precision  
Precision mengukur ketepatan model dalam memprediksi suatu kelas atau proporsi prediksi positif yang benar terhadap seluruh prediksi positif.
3. Recall  
Recall digunakan untuk mengukur kemampuan model dalam mengidentifikasi seluruh data yang termasuk ke dalam suatu kelas. Nilai recall yang tinggi menunjukkan kemampuan model dalam meminimalkan kesalahan *false negative*.
4. F1-score  
F1-score merupakan rata-rata harmonik antara precision dan recall, sehingga digunakan untuk menilai keseimbangan performa model terutama pada data yang tidak seimbang.

Semakin tinggi nilai evaluasi maka semakin baik performa model klasifikasi [10].

## 3. HASIL DAN PEMBAHASAN

Dataset prediksi tingkat obesitas digunakan pada penelitian ini. Dataset diperoleh dari platform Kaggle (*Obesity Levels Dataset*) dan digunakan untuk melakukan klasifikasi tingkat obesitas berdasarkan karakteristik fisik, pola makan, aktivitas harian, serta gaya hidup individu. Dataset terdiri dari 2.111 data dengan 16 atribut prediktor dan 1 atribut target yaitu *NObesyedad*. Variabel target terdiri dari tujuh kelas yaitu *Insufficient Weight*, *Normal Weight*, *Overweight Level I*, *Overweight Level II*, *Obesity Type I*, *Obesity Type II*, dan *Obesity Type III*. Atribut yang digunakan meliputi gender, age, height, weight, family\_history\_with\_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, dan MTRANS. Tabel 1 menunjukkan sebagian data penelitian yang digunakan.

**Tabel 2.** Dataset Penelitian

Age	Gender	Height	Weight	family_history_with_overweight	...	NObeyesdad
21	Female	1.62	64	yes	...	Normal Weight
23	Male	1.80	87	yes	...	Overweight I
19	Female	1.50	45	no	...	Insufficient Weight
28	Male	1.75	110	yes	...	Obesity Type I
...	...	...	...	...	...	...

### 3.1 Hasil Preprocessing Data

Tahap awal penelitian dilakukan melalui proses *preprocessing* untuk meningkatkan kualitas data sebelum dilakukan pemodelan. Berdasarkan hasil eksplorasi data menunjukkan bahwa dataset tidak memiliki *missing value* pada seluruh atribut sehingga semua data dapat digunakan dalam penelitian.

**Tabel 3.** *Missing Value* Dataset

No	Atribut	Missing Value	No	Atribut	Missing Value
1	Gender	0	9	CAEC	0
2	Age	0	10	CH2O	0
3	Height	0	11	FAF	0
4	Weight	0	12	TUE	0
5	family_history_with_overweight	0	13	CALC	0
6	FAVC	0	14	MTRANS	0
7	FCVC	0	15	NObeyesdad	0
8	NCP	0			

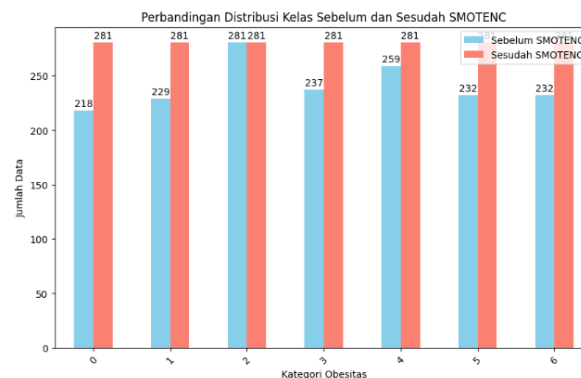
Berdasarkan Tabel 3 diketahui bahwa seluruh atribut memiliki nilai kosong sebesar 0 sehingga tidak dilakukan proses imputasi data. Selanjutnya dilakukan *label encoding* pada atribut kategorikal. Perbandingan data sebelum dan sesudah *label encoding* terdapat pada Tabel 4

**Tabel 4.** Perbandingan Data Sebelum dan Sesudah Label Encoding

No	Atribut	Sebelum Label Encoding	Sesudah Label Encoding	No	Atribut	Sebelum Label Encoding	Sesudah Label Encoding
1	Gender	Female	0	15	CAEC	Frequently	2
2	Gender	Male	1	16	CAEC	Always	3
3	CALC	no	0	17	MTRANS	Automobile	0
4	CALC	Sometimes	1	18	MTRANS	Bike	1
5	CALC	Frequently	2	19	MTRANS	Motorbike	2
6	CALC	Always	3	20	MTRANS	Public Transportation	3

7	FAVC	no	0	21	MTRANS	Walking	4
8	FAVC	yes	1	22	NObeyesdad	Insufficient Weight	0
9	SMOKE	no	0	23	NObeyesdad	Normal Weight	1
10	SMOKE	yes	1	24	NObeyesdad	Overweight Level I	2
11	family_history_with_overweight	no	0	25	NObeyesdad	Overweight Level II	3
12	family_history_with_overweight	yes	1	26	NObeyesdad	Obesity Type I	4
13	CAEC	no	0	27	NObeyesdad	Obesity Type II	5
14	CAEC	Sometimes	1	28	NObeyesdad	Obesity Type III	6

Kemudian dilakukan transformasi data numerik menggunakan StandardScaler untuk menyamakan skala antar atribut numerik sehingga setiap fitur memiliki rata-rata mendekati 0 dan standar deviasi sebesar 1. Proses ini bertujuan untuk mengurangi perbedaan rentang nilai antar fitur yang berpotensi memengaruhi proses pembelajaran model serta menjaga konsistensi tahapan preprocessing. Dataset kemudian dibagi menjadi 80% data training dan 20% data testing menggunakan stratified split agar proporsi kelas target tetap terjaga. Selanjutnya, SMOTENC diterapkan pada data training untuk menangani ketidakseimbangan distribusi kelas dengan mempertimbangkan karakteristik atribut numerik dan kategorikal secara bersamaan. Perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTENC ditunjukkan pada Gambar 2.



**Gambar 2.** Perbandingan Distribusi Kelas Sebelum dan Sesudah SMOTENC

Gambar 2 menunjukkan perbandingan jumlah data pada setiap kategori obesitas sebelum dan sesudah penerapan metode SMOTENC. Sebelum dilakukan SMOTENC, distribusi data terlihat tidak seimbang, di mana beberapa kelas memiliki jumlah data lebih sedikit dibandingkan kelas lainnya. Ketidakseimbangan ini berpotensi menyebabkan model cenderung bias terhadap kelas mayoritas. Setelah diterapkan SMOTENC, jumlah data pada setiap kategori menjadi sama (281 data per kelas). Hal ini menunjukkan bahwa SMOTENC berhasil menyeimbangkan distribusi data (*class balancing*) dengan cara menambah data sintesis pada kelas minoritas. Dengan distribusi yang lebih seimbang, model diharapkan dapat belajar secara lebih optimal dan tidak bias terhadap kelas tertentu.

### 3.2 Hasil Pengujian Model

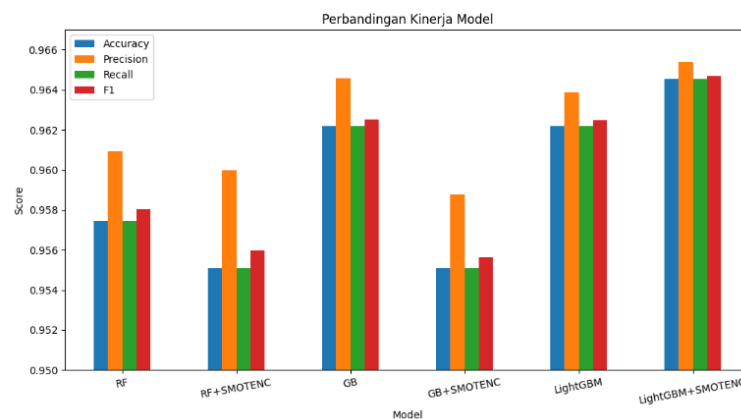
Setelah dilakukan proses pelatihan dan pengujian terhadap enam skenario model, selanjutnya dilakukan evaluasi performa menggunakan *metrik accuracy, precision, recall, dan F1-score*. Pengukuran ini bertujuan untuk mengetahui kemampuan masing-masing algoritma dalam melakukan klasifikasi tingkat obesitas serta menganalisis pengaruh penerapan SMOTENC terhadap peningkatan atau penurunan performa model. Hasil evaluasi dari seluruh model ditampilkan pada Tabel 5.

**Tabel 5.** Hasil Perbandingan Model

No	Model	Accuracy	Precision	Recall	F1-score
----	-------	----------	-----------	--------	----------

1	LightGBM + SMOTENC	0.9645	0.9654	0.9645	0.9647
2	Gradient Boosting	0.9622	0.9646	0.9622	0.9625
3	LightGBM	0.9622	0.9639	0.9622	0.9625
4	Random Forest	0.9574	0.9609	0.9574	0.9580
5	Random Forest + SMOTENC	0.9551	0.9600	0.9551	0.9560
6	Gradient Boosting + SMOTENC	0.9551	0.9588	0.9551	0.9556

Berdasarkan Tabel 5, model LightGBM dengan SMOTENC menunjukkan performa terbaik dibandingkan model lainnya dengan nilai accuracy sebesar 96,45%, precision 96,54%, recall 96,45%, dan F1-score 96,47%. Hasil ini menunjukkan bahwa SMOTENC mampu meningkatkan kemampuan LightGBM dalam menangani ketidakseimbangan data sehingga menghasilkan klasifikasi yang lebih optimal. Hal ini juga mengindikasikan bahwa LightGBM lebih adaptif terhadap data sintesis dibandingkan algoritma lainnya. Pada model tanpa SMOTENC, Gradient Boosting dan LightGBM memiliki nilai accuracy yang sama sebesar 96,22%, menunjukkan kemampuan klasifikasi yang tinggi meskipun tanpa penyeimbangan data. Sementara itu, Random Forest memperoleh accuracy sebesar 95,74%, yang relatif lebih rendah, sehingga mengindikasikan bahwa metode boosting lebih efektif dalam menangkap pola kompleks dibandingkan metode bagging. Namun, penerapan SMOTENC tidak selalu meningkatkan performa. Pada Random Forest, accuracy menurun menjadi 95,51%, dan pada Gradient Boosting juga mengalami penurunan menjadi 95,51%. Hal ini menunjukkan bahwa data sintesis tidak selalu meningkatkan generalisasi model. Secara keseluruhan, kombinasi LightGBM dan SMOTENC merupakan pendekatan terbaik dalam penelitian ini, sedangkan efektivitas SMOTENC pada algoritma lain bergantung pada karakteristik model yang digunakan.



**Gambar 3.** Perbandingan Kinerja Model

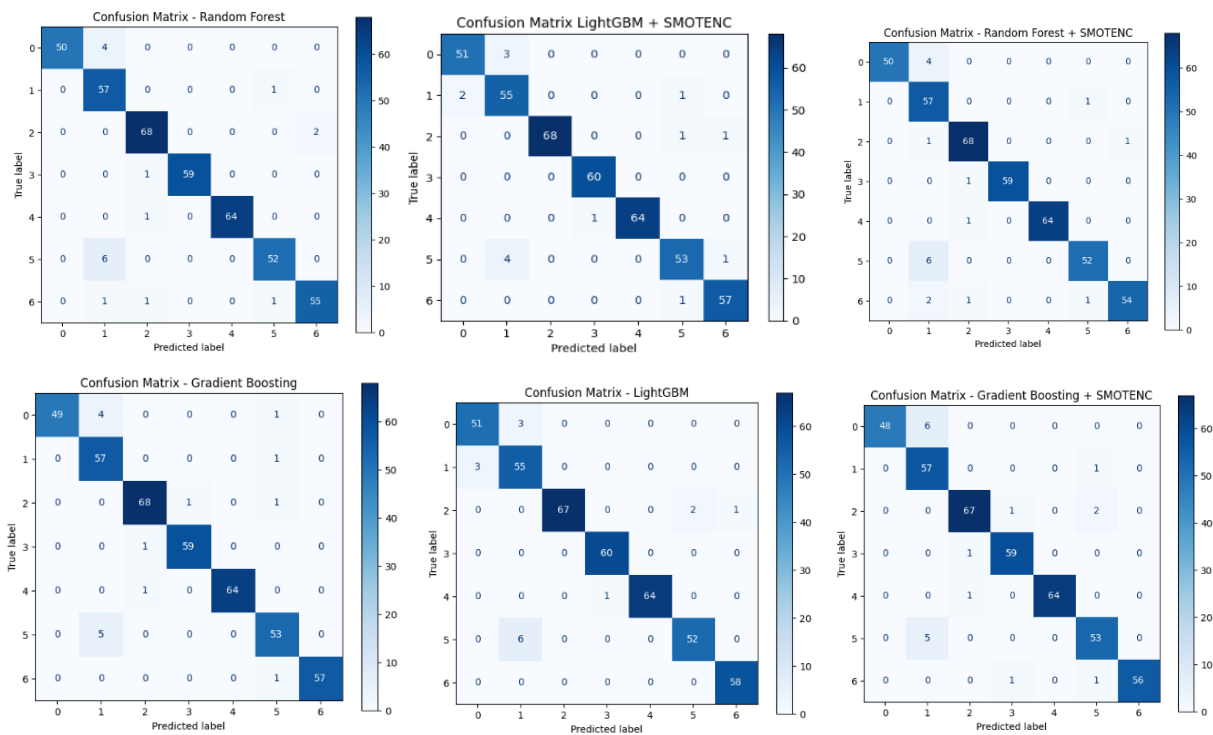
Pada perbandingan kinerja model klasifikasi. Secara umum, model yang digunakan, yaitu *Random Forest*, *Gradient Boosting*, dan *LightGBM*, menunjukkan performa yang tinggi pada seluruh metrik evaluasi (*accuracy*, *precision*, *recall*, dan *F1-score*). Model *LightGBM* dengan *SMOTENC* menghasilkan kinerja terbaik dibandingkan model lainnya, yang ditunjukkan oleh nilai metrik evaluasi tertinggi dan paling konsisten. Sementara itu, *Gradient Boosting* menunjukkan performa yang stabil baik sebelum maupun sesudah *SMOTENC*. Di sisi lain, *Random Forest* mengalami sedikit penurunan performa setelah penerapan *SMOTENC*, mengindikasikan bahwa efek oversampling tidak selalu berdampak positif pada semua algoritma. Secara keseluruhan, penerapan *SMOTENC* terbukti efektif dalam meningkatkan keseimbangan data dan berkontribusi pada peningkatan kualitas prediksi model, khususnya dalam kemampuan mengenali seluruh kelas secara lebih merata.

### 3.3 Analisis Pengaruh SMOTENC terhadap Model

Hasil pengujian menunjukkan bahwa penerapan teknik oversampling *SMOTENC* memberikan pengaruh yang berbeda terhadap performa masing-masing algoritma *ensemble learning* yang digunakan. Pada model *LightGBM*, penggunaan *SMOTENC* mampu meningkatkan nilai akurasi dari 96,21% menjadi 96,45%, disertai peningkatan pada metrik *precision*, *recall*, dan *F1-score*. Hasil ini mengindikasikan bahwa *LightGBM* memiliki kemampuan yang lebih baik dalam memanfaatkan distribusi data sintesis hasil oversampling sehingga mampu meningkatkan kemampuan klasifikasi terhadap kelas minoritas. Sebaliknya, penerapan *SMOTENC* pada model *Random Forest* menunjukkan penurunan performa dengan nilai akurasi dari 95,74% menjadi 95,50%, sedangkan pada *Gradient Boosting* terjadi penurunan akurasi dari 96,21% menjadi 95,50%. Penurunan tersebut menunjukkan bahwa penambahan data sintesis tidak selalu memberikan dampak positif terhadap seluruh algoritma, bahkan dapat menyebabkan model mengalami *overfitting*

atau kesulitan dalam mengenali pola asli data. Dengan demikian, hasil penelitian ini menegaskan bahwa efektivitas SMOTENC sangat dipengaruhi oleh karakteristik dataset serta kemampuan masing-masing algoritma dalam mempelajari distribusi data hasil *oversampling*. Berdasarkan keseluruhan hasil evaluasi, kombinasi LightGBM + SMOTENC menjadi model terbaik dengan nilai *accuracy* sebesar 96,45%, sehingga pendekatan tersebut dinilai paling optimal dalam menangani ketidakseimbangan kelas pada klasifikasi tingkat obesitas.

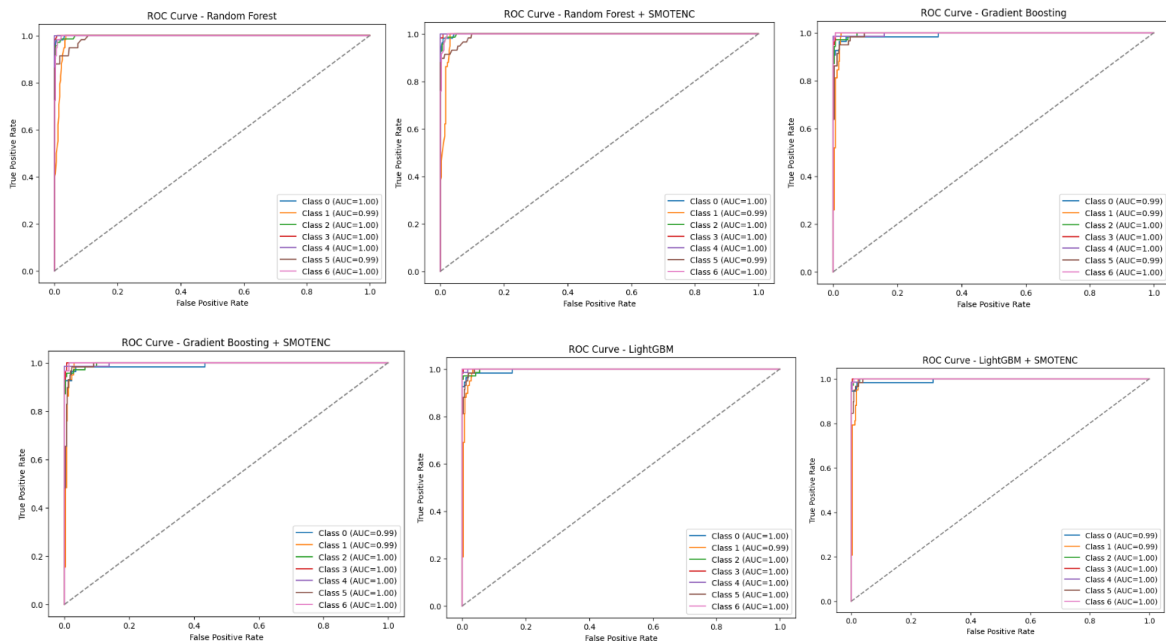
### 3.4 Confusion Matrik



**Gambar 4.** Confusion Matrik

Confusion matrix pada Gambar menunjukkan bahwa model Random Forest, Gradient Boosting, dan LightGBM memiliki performa klasifikasi yang baik, ditandai dengan dominasi nilai pada diagonal matriks yang merepresentasikan prediksi benar. Sebaliknya, nilai di luar diagonal menunjukkan kesalahan klasifikasi yang secara umum masih relatif kecil dan cenderung terjadi pada kelas-kelas yang berdekatan. Pada Random Forest, meskipun sebagian besar prediksi sudah tepat, masih terdapat kesalahan pada beberapa kelas, terutama kelas dengan jumlah data lebih sedikit. Setelah penerapan SMOTENC, distribusi prediksi menjadi lebih merata, namun peningkatan performa tidak terlalu signifikan. Sementara itu, Gradient Boosting menunjukkan kinerja yang stabil baik sebelum maupun sesudah SMOTENC, dengan sedikit perbaikan dalam mendeteksi kelas minoritas. Model LightGBM menunjukkan performa terbaik dibandingkan model lainnya, dengan nilai diagonal paling tinggi dan tingkat kesalahan paling rendah. Penerapan SMOTENC semakin meningkatkan kemampuan model ini dalam mengurangi kesalahan pada kelas minoritas dan mengasihkan distribusi prediksi yang lebih seimbang. Secara keseluruhan, SMOTENC terbukti membantu meningkatkan keseimbangan klasifikasi dan mengurangi bias terhadap kelas mayoritas, dengan hasil optimal pada model LightGBM.

### 3.5 Hasil ROC Curve



**Gambar 5. ROC Curve**

Berdasarkan hasil pengujian menggunakan kurva ROC (*Receiver Operating Characteristic*), seluruh model klasifikasi menunjukkan performa yang sangat baik dalam membedakan setiap kelas pada dataset. Nilai AUC (*Area Under Curve*) yang dihasilkan pada masing-masing model berada pada rentang 0,99 hingga 1,00, yang menandakan kemampuan klasifikasi yang sangat tinggi. Model *Random Forest* menghasilkan kurva ROC yang mendekati titik kiri atas grafik dengan nilai AUC hampir sempurna pada seluruh kelas. Setelah dilakukan penyeimbangan data menggunakan metode SMOTENC, performa model mengalami peningkatan stabilitas dengan nilai AUC tetap tinggi pada setiap kelas. Pada model *Gradient Boosting*, kurva ROC menunjukkan performa klasifikasi yang sangat optimal dengan rata-rata nilai AUC sebesar 0,99 hingga 1,00. Penggunaan SMOTENC pada model ini juga membantu meningkatkan kemampuan model dalam mengenali kelas minoritas sehingga distribusi prediksi menjadi lebih seimbang. Sementara itu, model *LightGBM* memperlihatkan hasil terbaik dengan kurva ROC yang sangat dekat dengan nilai maksimum dan AUC mencapai 1,00 pada beberapa kelas. Setelah diterapkan SMOTENC, performa model tetap konsisten dan mampu mempertahankan tingkat sensitivitas serta spesifisitas yang tinggi. Secara keseluruhan, hasil ROC Curve menunjukkan bahwa ketiga algoritma memiliki kemampuan klasifikasi yang sangat baik dalam memprediksi data penelitian. Nilai AUC yang mendekati 1 menunjukkan bahwa model mampu membedakan antar kelas secara akurat. Selain itu, penerapan metode SMOTENC terbukti membantu meningkatkan keseimbangan data tanpa menurunkan performa model klasifikasi.

## 4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *ensemble learning*, yaitu *Random Forest*, *Gradient Boosting*, dan *LightGBM*, menunjukkan kinerja yang baik dalam melakukan klasifikasi tingkat obesitas. Hal ini dibuktikan melalui nilai metrik evaluasi yang tinggi, meliputi *accuracy*, *precision*, *recall*, dan *F1-score*, serta diperkuat oleh analisis *confusion matrix* yang menunjukkan dominasi nilai pada diagonal matriks sebagai indikator tingginya jumlah prediksi yang benar. Selain itu, kurva ROC juga mengindikasikan kemampuan diskriminatif model yang baik dalam membedakan antar kelas. Penerapan metode SMOTENC terbukti efektif dalam mengatasi permasalahan ketidakseimbangan data dengan meningkatkan distribusi kelas minoritas. Teknik ini mampu mengurangi bias model terhadap kelas mayoritas dan meningkatkan kemampuan generalisasi model dalam mengenali seluruh kelas. Meskipun demikian, pengaruh penerapan SMOTENC terhadap peningkatan performa model tidak bersifat seragam pada setiap algoritma. Pada model *Random Forest* dan *Gradient Boosting*, peningkatan performa yang dihasilkan cenderung terbatas, sedangkan pada model *LightGBM* terjadi peningkatan yang lebih signifikan. Model *LightGBM* dengan penerapan SMOTENC menghasilkan performa terbaik dibandingkan model lainnya, ditunjukkan oleh tingkat akurasi tertinggi serta kesalahan klasifikasi yang paling rendah. Dengan demikian, dapat disimpulkan bahwa kombinasi *algoritma boosting* dan teknik penyeimbangan data SMOTENC merupakan pendekatan yang efektif dalam meningkatkan kinerja klasifikasi pada dataset multikelas yang tidak seimbang. Temuan ini memberikan kontribusi dalam pengembangan model klasifikasi yang lebih optimal, khususnya pada bidang kesehatan.

## REFERENCES

- [1] L. Bastida *et al.*, "Promoting Obesity Prevention and Healthy Habits in Childhood: The OCARIoT Experience," *IEEE J. Transl. Eng. Heal. Med.*, vol. 11, no. February, pp. 261–270, 2023, doi: 10.1109/JTEHM.2023.3261899.
- [2] M. E. Piché, A. Tchernof, and J. P. Després, "Obesity Phenotypes, Diabetes, and Cardiovascular Diseases," *Circ. Res.*, vol. 126, no. 11, pp. 1477–1500, 2020, doi: 10.1161/CIRCRESAHA.120.316101.
- [3] M. Azad, M. F. K. Khan, and S. A. El-Ghany, "XAI-Enhanced Machine Learning for Obesity Risk Classification: A Stacking Approach With LIME Explanations," *IEEE Access*, vol. 13, no. December 2024, pp. 13847–13865, 2025, doi: 10.1109/ACCESS.2025.3530840.
- [4] A. Salam and N. Abhinesh, "Revolutionizing dermatology: The role of artificial intelligence in clinical practice," *IP Indian J. Clin. Exp. Dermatology*, vol. 10, no. 2, pp. 107–112, 2024, doi: 10.18231/j.ijced.2024.021.
- [5] B. Wang *et al.*, "Diastolic and systolic right ventricular diameters for predicting pulmonary hypertension in children with congenital heart disease," *Clin. Imaging*, vol. 70, no. March 2020, pp. 67–73, 2021, doi: 10.1016/j.clinimag.2020.10.027.
- [6] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, no. Nips, pp. 3147–3155, 2017.
- [7] M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, "Automatic Detection of Students' Engagement During Online Learning: A Bagging Ensemble Deep Learning Approach," *IEEE Access*, vol. 12, no. June, pp. 96063–96073, 2024, doi: 10.1109/ACCESS.2024.3425820.
- [8] M. Mukherjee and M. Khushi, "Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features," *Appl. Syst. Innov.*, vol. 4, no. 1, pp. 1–15, 2021, doi: 10.3390/asi4010018.
- [9] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [10] S. Yin, Z. Li, D. Song, H. Mu, Y. Niu, and X. Wang, "Study on the Energy Evolution Law and Bursting Liability of Coal Failure with Different Joint Inclination Angles," *Appl. Sci.*, vol. 14, no. 3, 2024, doi: 10.3390/app14031120.
- [11] H. Javed, H. A. Muqet, T. Javed, A. U. Rehman, and R. Sadiq, "Ethical Frameworks for Machine Learning in Sensitive Healthcare Applications," *IEEE Access*, vol. 12, no. October 2023, pp. 16233–16254, 2024, doi: 10.1109/ACCESS.2023.3340884.
- [12] B. F. Wu, Y. C. Wu, L. W. Chiu, and H. P. Liu, "Soft Label With Channel Encoding for Dependent Facial Image Classification," *IEEE Access*, vol. 10, pp. 10661–10672, 2022, doi: 10.1109/ACCESS.2022.3145195.
- [13] S. Ouaari, A. B. Ünal, M. Akgün, and N. Pfeifer, "Robust Representation Learning for Privacy-Preserving Machine Learning: A Multi-Objective Autoencoder Approach," *IEEE Access*, vol. 13, no. July, pp. 151527–151537, 2025, doi: 10.1109/ACCESS.2025.3603429.
- [14] A. Tharwat and W. Schenck, "Active Learning for Handling Missing Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 2, pp. 3273–3287, 2025, doi: 10.1109/TNNLS.2024.3352279.
- [15] A. S. Desuky, Y. M. Elbarawy, S. Kausar, A. H. Omar, and S. Hussain, "Single-Point Crossover and Jellyfish Optimization for Handling Imbalanced Data Classification Problem," *IEEE Access*, vol. 10, no. 3, pp. 11730–11749, 2022, doi: 10.1109/ACCESS.2022.3146424.
- [16] D. Muller, I. Soto-Rey, and F. Kramer, "An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks," *IEEE Access*, vol. 10, no. June, pp. 66467–66480, 2022, doi: 10.1109/ACCESS.2022.3182399.
- [17] E. Santi *et al.*, "High-Resolution Mapping of Soil Moisture by AMSR2 Data Disaggregation Based on Sentinel-1 and Machine Learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 15077–15088, 2024, doi: 10.1109/JSTARS.2024.3445111.
- [18] Y. Gao, L. Wang, G. Zhong, Y. Wang, and J. Yang, "Potential of Remote Sensing Images for Soil Moisture Retrieving Using Ensemble Learning Methods in Vegetation-Covered Area," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 8149–8165, 2023, doi: 10.1109/JSTARS.2023.3311096.
- [19] J. Erbani, P. É. Portier, E. Egyed-Zsigmond, and D. Nurbakova, "Confusion Matrices: A Unified Theory," *IEEE Access*, vol. 12, no. October, pp. 181372–181419, 2024, doi: 10.1109/ACCESS.2024.3507199.