Prosiding Seminar Nasional Indonesia
Vol. 4 No. 1 2026, hal., 61-76
https://sociohum.net/index.php/PROSIDINGNASIOANAL
e-ISSN: 3026-5169

KLASIFIKASI RISIKO OBESITAS BERBASIS GRADIENT BOOSTING PADA DATA MEDIS

Wildan Hafiz Firmansyah
Universitas Pembangunan Nasional "Veteran" Jawa Timur
22082010198@student.upnjatim.ac.id

Nala Widyadhana
Universitas Pembangunan Nasional "Veteran" Jawa Timur
22082010195@student.upnjatim.ac.id

Abstract

Obesity is a growing health concern that can lead to various chronic diseases, making accurate risk identification an important preventive effort. The development of machine learning techniques enables the utilization of medical data to support intelligent decision-making in the healthcare domain. This study aims to apply the Gradient Boosting algorithm as a classification method to predict obesity risk based on medical data. The dataset used contains information related to eating habits, physical activities, and individual characteristics. The research process includes data preprocessing, data transformation and normalization, class mapping, and data partitioning into training and testing sets with a ratio of 70:30. The Gradient Boosting model is constructed using multiple decision trees with specific parameter settings to classify obesity risk into two categories, namely obese and non-obese. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics. The experimental results show that the proposed model achieves good classification performance with an accuracy exceeding 90%, while the performance gap between training and testing data remains relatively small. This indicates that the model has strong generalization capability and does not suffer from overfitting. Therefore, the application of Gradient Boosting on medical data proves to be an effective approach for obesity risk classification and has the potential to support intelligent health information systems in assisting medical practitioners with more precise obesity prevention and management strategies.

Keywords: *gradient boosting, classification, obesity risk, medical data, machine learning.*

Abstrak

Obesitas merupakan permasalahan kesehatan yang terus meningkat dan berpotensi menimbulkan berbagai penyakit kronis, sehingga diperlukan upaya deteksi risiko secara akurat dan berbasis data. Perkembangan machine learning memberikan peluang dalam mengolah data medis untuk mendukung pengambilan keputusan di bidang kesehatan. Penelitian ini bertujuan menerapkan algoritma Gradient Boosting sebagai metode klasifikasi untuk memprediksi risiko obesitas berdasarkan data medis. Data yang digunakan berasal dari Obesity Dataset yang memuat informasi kebiasaan makan, aktivitas fisik, serta karakteristik individu. Tahapan penelitian meliputi prapemrosesan data,

transformasi dan normalisasi data, pemetaan kelas, serta pembagian data menjadi data latih dan data uji dengan rasio 70:30. Model Gradient Boosting dibangun menggunakan beberapa pohon keputusan dengan parameter tertentu untuk menghasilkan klasifikasi risiko obesitas menjadi dua kelas, yaitu obesitas dan tidak obesitas. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model Gradient Boosting mampu memberikan kinerja klasifikasi yang baik dengan tingkat akurasi lebih dari 90% serta perbedaan performa antara data latih dan data uji yang relatif kecil. Hal ini mengindikasikan bahwa model yang dibangun memiliki kemampuan generalisasi yang baik dan tidak mengalami overfitting. Dengan demikian, penerapan Gradient Boosting pada data medis dapat menjadi pendekatan yang efektif dalam mengklasifikasikan risiko obesitas dan berpotensi mendukung sistem informasi kesehatan dalam membantu tenaga medis melakukan pencegahan dan penanganan obesitas secara lebih tepat.
Kata Kunci : gradient boosting, klasifikasi, risiko obesitas, data medis, machine learning.

PENDAHULUAN

Obesitas merupakan permasalahan kesehatan yang prevalensinya terus meningkat dan berdampak signifikan terhadap kondisi fisik maupun kualitas hidup individu [5], [9]. Peningkatan risiko penyakit kronis, seperti diabetes melitus, gangguan kardiovaskular, dan penyakit metabolik lainnya, menjadikan obesitas sebagai isu yang membutuhkan perhatian khusus dalam upaya pencegahan dan penanganan kesehatan masyarakat. Oleh karena itu, diperlukan pendekatan yang mampu mengidentifikasi risiko obesitas secara lebih akurat dan sistematis.

Perkembangan teknologi informasi mendorong pemanfaatan data medis dalam sistem informasi kesehatan sebagai salah satu solusi yang relevan. Data yang mencerminkan kebiasaan makan, aktivitas fisik, serta karakteristik individu dapat diolah untuk menghasilkan informasi yang bernilai dalam mendukung pengambilan keputusan [14]. Dalam konteks ini, machine learning berperan sebagai metode analitik yang mampu mengenali pola tersembunyi dalam data dan membangun model prediktif berdasarkan data historis [6], [10], [12].

Metode klasifikasi dalam machine learning memungkinkan pengelompokan individu ke dalam kategori risiko tertentu berdasarkan karakteristik yang dimiliki [15]. Salah satu algoritma yang banyak digunakan adalah Gradient Boosting, yaitu metode ensemble yang mengkombinasikan beberapa model prediktor sederhana untuk menghasilkan model yang lebih akurat [2], [11], [13]. Penelitian ini bertujuan menerapkan algoritma Gradient Boosting dalam mengklasifikasikan risiko obesitas berdasarkan data medis melalui tahapan prapemrosesan data, pembangunan model, dan evaluasi kinerja. Hasil penelitian diharapkan dapat berkontribusi dalam pengembangan sistem informasi kesehatan berbasis kecerdasan buatan sebagai alat bantu tenaga medis dalam upaya pencegahan dan penanganan obesitas berbasis data.

METODE PENELITIAN

Metodologi penelitian ini menjelaskan secara rinci langkah-langkah yang dilakukan untuk mencapai tujuan penelitian, yaitu mengklasifikasikan risiko obesitas menggunakan algoritma *Gradient Boosting* pada data medis. Bagian ini mencakup deskripsi dataset, teknik pengumpulan data, proses analisis data, tahapan *preprocessing*, pemodelan, serta evaluasi hasil. Dengan penjelasan yang terperinci, diharapkan pembaca dapat memahami dan mereplikasi prosedur penelitian ini dengan hasil yang serupa.

2.1 Data

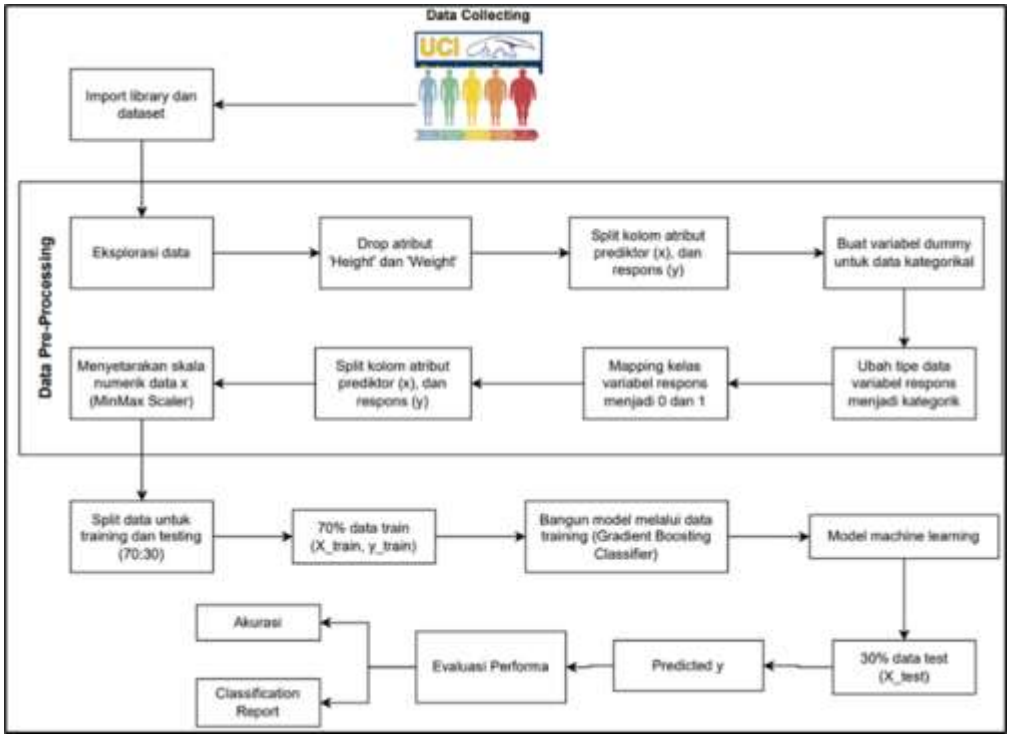
Data yang digunakan dalam penelitian ini merupakan dataset publik yang diambil dari UCI Machine Learning Repository dengan judul *"Obesity Levels based on Eating Habits and Physical Condition."* [7]. Dataset ini berisi data individu dari tiga negara, yaitu Meksiko, Peru, dan Kolombia, dengan rentang usia antara 14 hingga 61 tahun. Dataset terdiri dari 2.111 baris data (records) dan 17 atribut (kolom) yang menggambarkan kebiasaan makan, aktivitas fisik, serta kondisi fisik responden. Data tersebut memiliki kelas seimbang (balanced), sehingga distribusi kategori antara obesitas dan tidak obesitas relatif merata. Atribut-atribut dalam dataset meliputi:

- Atribut kebiasaan makan: FAVC (frekuensi konsumsi makanan berkalori tinggi), FCVC (frekuensi konsumsi sayuran), NCP (jumlah makanan utama), CAEC (kebiasaan makan di antara waktu makan), CH2O (konsumsi air harian), dan CALC (konsumsi alkohol).
- Atribut kondisi fisik: SCC (pemantauan konsumsi kalori), FAF (frekuensi aktivitas fisik), TUE (durasi penggunaan perangkat teknologi), dan MTRANS (jenis transportasi yang digunakan).
- Atribut demografis: Gender, Age, Height, dan Weight.
- Variabel target: *NObesity* yang menunjukkan tingkat obesitas individu.

2.2 Teknik Pengumpulan Data

Data dikumpulkan dari sumber resmi UCI Machine Learning Repository, yang telah terverifikasi dan digunakan secara luas dalam penelitian *machine learning* [7].
Selain itu, dilakukan studi pustaka dengan mengacu pada buku, jurnal ilmiah, serta publikasi yang relevan untuk memperkuat teori dan pemilihan metode penelitian.
Pendekatan ini memastikan bahwa proses penelitian memiliki dasar ilmiah yang kuat serta memungkinkan replikasi oleh peneliti lain.

2.3 Teknik Analisis Data



Gambar 1. Alur Proses Analisis

Berdasarkan Gambar 1 analisis data dilakukan melalui tahapan sistematis yang melibatkan data preprocessing, pemisahan data, pemodelan dengan algoritma *Gradient Boosting*, dan evaluasi performa model. Tahapan ini dirancang agar penelitian dapat diulangi oleh peneliti lain dengan hasil yang sama.

2.3.1 Data Preprocessing

Tahapan *preprocessing* dilakukan untuk memastikan bahwa data yang digunakan memiliki kualitas baik dan siap untuk dianalisis menggunakan algoritma *machine learning* [1], [6]. Langkah-langkah yang dilakukan meliputi:

1. Pembersihan Data (Data Cleaning)
   Menghapus nilai kosong (*missing values*) dan memastikan setiap atribut memiliki format data yang sesuai.
2. Integrasi dan Transformasi Data
   Menggabungkan seluruh atribut ke dalam satu dataset terstruktur, kemudian melakukan transformasi data kategorik menjadi numerik menggunakan metode *One-Hot Encoding*. Contohnya, atribut seperti "*Gender*" diubah menjadi nilai biner (0 = laki-laki, 1 = perempuan).
3. Normalisasi Data
   Data numerik dinormalisasi menggunakan metode *Min-Max Scaler*, sehingga semua nilai berada dalam rentang 0–1. Proses ini membantu model *Gradient*

Boosting bekerja lebih optimal dan menghindari dominasi fitur dengan nilai besar.

4. Mapping Kelas (*Reclassifying Labels*)

Variabel target *NObesity* awalnya memiliki tujuh kelas, yaitu:

- Normal_Weight
- Overweight_Level
- Overweight_Level
- Insufficient_Weigh
- Obesity_Type_I
- Obesity_Type_II
- Obesity_Type_II

Dalam penelitian ini, kelas tersebut digabung menjadi dua kategori utama:

- 0 = Tidak Obesitas (Insufficient Weight, Normal, Overweight I, Overweight II)
- 1 = Obesitas (Obesity Type I, II, dan III).

Langkah ini dilakukan agar model lebih fokus pada klasifikasi risiko obesitas secara biner.

2.3.2 Splitting Dataset

Setelah dilakukan *preprocessing*, dataset dibagi menjadi dua bagian:

1. Data latih (*training set*) sebesar 70% dari total data,
2. Data uji (*testing set*) sebesar 30% dari total data.

Pembagian ini dilakukan menggunakan fungsi *train_test_split* dari pustaka *Scikit-Learn*. Data latih digunakan untuk melatih model, sementara data uji digunakan untuk mengevaluasi kemampuan model dalam mengenali data baru yang belum pernah dilihat sebelumnya.

2.3.3 Pemodelan dengan Gradient Boosting

Proses pemodelan dilakukan menggunakan algoritma *Gradient Boosting Classifier* yang tersedia dalam pustaka *Scikit-Learn*. Algoritma ini bekerja dengan membangun model secara bertahap, di mana setiap model baru memperbaiki kesalahan dari model sebelumnya [2], [11]. Parameter utama yang digunakan dalam penelitian ini meliputi:

- ***n_estimators* = 100** (jumlah pohon yang dibentuk),
- ***learning_rate* = 0.1** (kecepatan pembelajaran model),
- ***max_depth* = 4** (kedalaman maksimum setiap pohon),
- ***random_state* = 0** (untuk menjaga konsistensi hasil).

Model dilatih menggunakan data latih, kemudian dilakukan pengujian dengan data uji untuk menghasilkan prediksi tingkat risiko obesitas.

2.3.4 Evaluasi Model

Evaluasi model dilakukan dengan mengukur tingkat akurasi dan efektivitas algoritma dalam melakukan klasifikasi [3], [8]. Beberapa metrik yang digunakan antara lain:

- **Akurasi (*Accuracy*):** Mengukur persentase prediksi yang benar dibandingkan total data uji.
- **Presisi (*Precision*):** Menunjukkan ketepatan model dalam memprediksi kelas positif.
- **Recall (Sensitivitas):** Mengukur seberapa baik model dalam mengenali data yang benar-benar positif.
- **F1-Score:** Kombinasi antara presisi dan *recall* untuk memberikan evaluasi yang seimbang.

Hasil evaluasi menunjukkan bahwa model *Gradient Boosting* mampu mencapai akurasi sebesar 90,38%, yang berarti model dapat memprediksi risiko obesitas dengan tingkat ketepatan yang tinggi tanpa mengalami *overfitting* maupun *underfitting*.

HASIL DAN PEMBAHASAN

Setelah menentukan metodologi, peneliti melanjutkan tahapan penelitian dengan melakukan pengolahan dan analisis data secara sistematis untuk memperoleh hasil yang sesuai dengan tujuan penelitian. Pada bab ini dipaparkan hasil dari setiap tahap yang telah dilakukan, mulai dari preprocessing data hingga tahap evaluasi model.

3.1 Deskripsi Data

Dataset risiko obesitas yang digunakan dalam penelitian ini merupakan kumpulan data numerik dan kategorik yang berkaitan dengan tingkat obesitas individu [7]. Tabel 1 menyajikan deskripsi variabel serta jenis nilai (value) yang terdapat dalam dataset obesitas.

Tabel 1. Variabel Dataset Obesitas

Variabel Respon		
NObesity	Kategorikal	Kurus, Normal, Gemuk_I, Gemuk_II, Gemuk_III, Obesitas_I, Obesitas_II, Obesitas_III
Variabel Prediktor		
Gender	Kategorikal	Female, Male
Age	Numerikal	Umur dengan angka numerik
Height	Numerikal	Tinggi badan dalam meter
Weight	Numerikal	Berat badan dalam kilogram
Family History with Overweight	Kategorikal	Riwayat obesitas keluarga (Ya/Tidak)
FAVC	Kategorikal	Makanan berkalori tinggi (Ya/Tidak)
FCVC	Numerikal	Konsumsi buah (Tidak Pernah/Terkadang/Selalu)
NCP	Numerikal	Frekuensi camilan (1-2 kali/3 kali/Lebih dari 3 kali)

CAEC	Kategorikal	Makanan pendamping (Tidak Pernah/Terkadang/ Sering/Selalu)
SMOKE	Kategorikal	Perokok (Ya/Tidak)
CH2O	Numerikal	Frekuensi minum air putih (<1 liter/1-2 liter/ >2 liter)
SCC	Kategorikal	Monitor kalori (Ya/Tidak)
FAF	Numerikal	Aktivitas fisik (Tidak Pernah/1-2 hari/2-4 hari/4-5 hari)
TUE	Numerikal	Penggunaan gadget (0-2 jam/3-5 jam/ > 5 jam)
CALC	Kategorikal	Konsumsi alkohol (Tidak Pernah/Terkadang/Sering)
MTRANS	Kategorikal	Transportasi (Mobil/Motor/Sepeda/Transportasi Umum/Berjalan)

Variabel *NObesity* merupakan variabel respon dengan tipe data kategorikal yang berperan sebagai **y** dalam proses klasifikasi machine learning. Sementara itu, 16 variabel lainnya berperan sebagai variabel prediktor (**x**), yang terdiri dari 8 variabel numerikal dan 8 variabel kategorikal.

3.2 Eksplorasi Variabel Numerik

Gambar 2. Korelasi antar variabel numerik

Berdasarkan Gambar 2 heatmap korelasi menunjukkan bahwa sebagian besar variabel numerik memiliki hubungan yang lemah. Korelasi paling menonjol terdapat antara Height dan Weight dengan hubungan positif sedang, sementara Age berkorelasi negatif lemah dengan TUE dan FAF. Variabel lainnya umumnya memiliki nilai korelasi mendekati nol, sehingga tidak menunjukkan hubungan linear yang kuat.

3.3 Data Preprocessing

1. Seleksi Variabel

	Height	Weight	NObeyesdad	BMI	BMI_Label
0	1.62	64.00	Normal_Weight	24.39	Normal
1	1.52	56.00	Normal_Weight	24.24	Normal
2	1.80	77.00	Normal_Weight	23.77	Normal
3	1.80	87.00	Overweight_Level_I	26.85	Overweight
4	1.78	89.80	Overweight_Level_II	28.34	Overweight
...	...	...	...	...	...
2106	1.71	131.41	Obesity_Type_III	44.90	Obesitas III
2107	1.75	133.74	Obesity_Type_III	43.74	Obesitas III
2108	1.75	133.69	Obesity_Type_III	43.54	Obesitas III
2109	1.74	133.35	Obesity_Type_III	44.07	Obesitas III
2110	1.74	133.47	Obesity_Type_III	44.14	Obesitas III

Gambar 3. Perhitungan label obesitas dengan BMI

Berdasarkan Gambar 3 kategori obesitas ditentukan berdasarkan perhitungan Indeks Massa Tubuh (BMI) yang berasal dari variabel *weight* dan *height*. Oleh karena itu, kedua variabel tersebut dihilangkan untuk mencegah *data leakage* dan memastikan proses klasifikasi berfokus pada pengaruh variabel prediktor lainnya.

2. Mapping Kelas Data

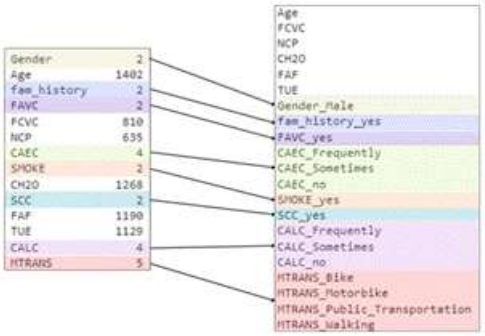
Variabel respons pada dataset obesitas memiliki 7 kelas. Dari 7 kategori yang ada kemudian akan dibagi kembali menjadi dua kelas, yaitu obesitas (1) dan tidak obesitas (0) seperti yang terlihat pada Gambar 4.

	Gender	Age	Height	Weight	fam_history	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad_y
0	Female	21.00	1.62	64.00	yes	no	2.00	3.00	Sometimes	no	2.00	no	0.00	1.00	no	Public_Transportation	0
1	Female	21.00	1.52	56.00	yes	no	3.00	3.00	Sometimes	yes	3.00	yes	3.00	0.00	Sometimes	Public_Transportation	0
2	Male	23.00	1.80	77.00	yes	no	2.00	3.00	Sometimes	no	2.00	no	2.00	1.00	Frequently	Public_Transportation	0
3	Male	27.00	1.80	87.00	no	no	3.00	3.00	Sometimes	no	2.00	no	2.00	0.00	Frequently	Walking	0
4	Male	22.00	1.78	89.80	no	no	2.00	1.00	Sometimes	no	2.00	no	0.00	0.00	Sometimes	Public_Transportation	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	Female	20.98	1.71	131.41	yes	yes	3.00	3.00	Sometimes	no	1.73	no	1.68	0.91	Sometimes	Public_Transportation	1
2107	Female	21.98	1.75	133.74	yes	yes	3.00	3.00	Sometimes	no	2.01	no	1.34	0.60	Sometimes	Public_Transportation	1
2108	Female	22.52	1.75	133.69	yes	yes	3.00	3.00	Sometimes	no	2.05	no	1.41	0.65	Sometimes	Public_Transportation	1
2109	Female	24.36	1.74	133.35	yes	yes	3.00	3.00	Sometimes	no	2.85	no	1.14	0.59	Sometimes	Public_Transportation	1
2110	Female	23.66	1.74	133.47	yes	yes	3.00	3.00	Sometimes	no	2.86	no	1.03	0.71	Sometimes	Public_Transportation	1

2111 rows × 17 columns

Gambar 4. Mapping Kelas Data

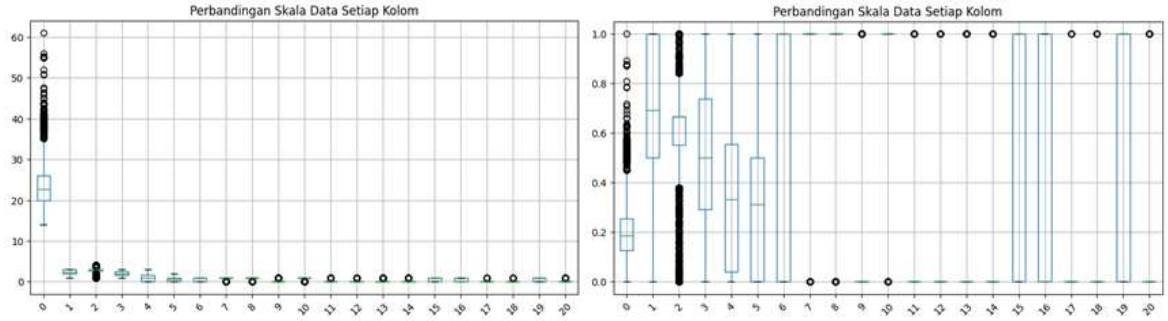
3. Transformasi Data



Gambar 5. Transformasi variabel dummy

Berdasarkan Gambar 5 transformasi data dilakukan dengan encoding untuk mengubah data kategorik menjadi numerik agar dapat diproses oleh machine learning. Pada penelitian ini digunakan one-hot encoding pada variabel prediktor (x), sehingga data kategorik diubah menjadi variabel dummy dengan nilai 0 dan 1.

4. Normalisasi Data



Gambar 6. Boxplot sebelum(kiri) dan sesudah (kanan) tahap normalisasi

Pada Gambar 6 perbedaan skala antar atribut data memerlukan proses normalisasi sebelum pemodelan machine learning. Normalisasi dilakukan menggunakan Min-Max Scaler dari Scikit-Learn untuk menyamakan skala data ke rentang 0–1, sehingga kinerja model menjadi lebih optimal dan stabil.

5. Splitting Dataset

Data disiapkan untuk pemodelan dengan memisahkan variabel prediktor (x) dan variabel respons (y), kemudian dibagi menjadi data latih dan data uji. Pembagian data dilakukan dengan rasio 70:30, yaitu 70% sebagai data training dan 30% sebagai data testing.

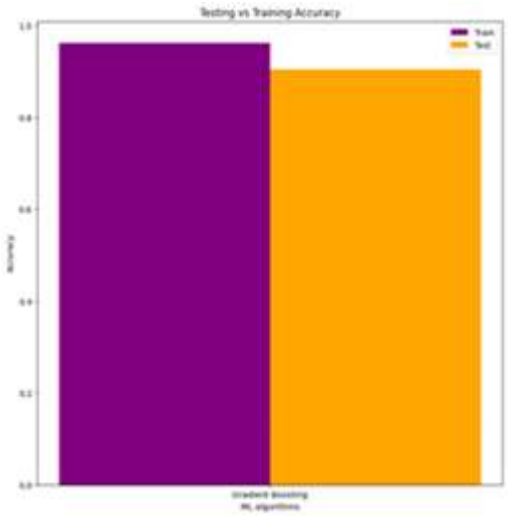
3.4 Klasifikasi Gradient Boosting

Proses klasifikasi pada penelitian ini dilakukan menggunakan algoritma Gradient Boosting Classifier dari pustaka Scikit-Learn pada bahasa pemrograman Python. Model dibangun dengan memanfaatkan data latih (training set) dan parameter n_estimators = 100, learning_rate = 0.1, max_depth = 4, dan random_state = 0.

	precision	recall	f1-score	support
0	0.93	0.89	0.91	340
1	0.88	0.92	0.90	294
accuracy			0.90	634
macro avg	0.90	0.90	0.90	634
weighted avg	0.90	0.90	0.90	634

Gambar 7. Akurasi model

Berdasarkan Gambar 7, hasil pelatihan model menghasilkan nilai akurasi sebesar 0,91 pada data latih dan 0,9038 atau sekitar 90,38% pada data uji. Selisih akurasi antara data latih dan data uji relatif kecil, sehingga dapat disimpulkan bahwa model memiliki kinerja yang baik dan tidak mengalami kondisi *overfitting* maupun *underfitting*. Perbandingan akurasi data latih dan data uji ditunjukkan pada Gambar 8.



Gambar 8. Perbandingan akurasi data latih dan data uji

Dari hasil prediksi di atas, didapatkan nilai evaluasi performa yang bisa diperiksa dalam Classification Report beserta Confusion Matrix dari model.

1. Classification Report
   - Precision: 93% prediksi model untuk kelas 0 adalah benar positif (TP). Sedangkan sekitar 88% prediksi model untuk kelas 1 adalah benar positif.

$$Presisi\ 0 = \frac{TN}{(TN + FN)} = \frac{303}{(303 + 24)} = \frac{303}{327} \approx 0.9265$$

$$Presisi\ 1 = \frac{TP}{(TP + FP)} = \frac{270}{(270 + 37)} = \frac{270}{307} \approx 0.8799$$

- Recall: Model dapat mengenali sekitar 89% dari data yang sebenarnya positif untuk kelas 0. Dan model dapat mengenali sekitar 92% dari data yang sebenarnya positif untuk kelas 1.

$$Recall\ 0 = \frac{TN}{(TN + FP)} = \frac{303}{(303 + 37)} = \frac{303}{340} \approx 0.8912$$
$$Recall\ 1 = \frac{TP}{(TP + FN)} = \frac{270}{(270 + 24)} = \frac{270}{294} \approx 0.9184$$

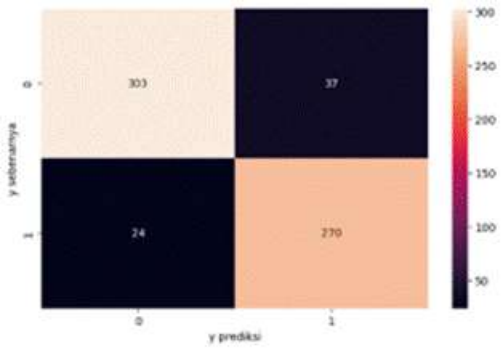
- F1-score: Dihitung dengan menggunakan harmonic mean dari precision dan recall. Dalam kasus ini, f1-score untuk kelas 0 adalah 0.91, dan untuk kelas 1 adalah 0.90.

$$F1\ Score = \frac{2 \times Presisi \times Recall}{Presisi + Recall} = \frac{2 \times 0.9265 \times 0.8912}{0.9265 \times 0.8912} \approx 0.9069$$
$$F1\ Score = \frac{2 \times Presisi \times Recall}{Presisi + Recall} = \frac{2 \times 0.8799 \times 0.9184}{0.8799 + 0.9184} \approx 0.8985$$

- Support: Terdapat 340 sampel dalam kelas 0, dan terdapat 294 sampel dalam kelas 1.

$$Support\ kelas\ 0 = TN + FP = 303 + 37 = 340$$
$$Support\ kelas\ 1 = FN + TP = 24 + 270 = 294$$

2. Confusion Matrix:



Gambar 9. Confusion Matrix

True Negative (TN)	:	Terdapat 303 data yang sebenarnya negatif dan diprediksi benar sebagai negatif.
False Positive (FP)	:	Terdapat 37 data yang sebenarnya negatif tetapi diprediksi salah sebagai positif.
False Negative (FN)	:	Terdapat 24 data yang sebenarnya positif tetapi diprediksi salah sebagai negatif.

True Positive (TP)	:	Terdapat 270 data yang sebenarnya positif dan diprediksi benar sebagai positif.

Analisis/Diskusi

Hasil penelitian menunjukkan bahwa algoritma Gradient Boosting Classifier mampu memberikan performa klasifikasi yang baik dalam memprediksi risiko obesitas berdasarkan data medis. Model yang dibangun menghasilkan tingkat akurasi sebesar 90,38% pada data uji dengan selisih performa yang relatif kecil dibandingkan data latih. Selisih akurasi antara data latih sebesar 91% dan data uji sebesar 90,38% menunjukkan bahwa model memiliki kemampuan generalisasi yang baik dan tidak mengalami kondisi overfitting maupun underfitting. Kondisi ini mengindikasikan bahwa model mampu mempelajari pola data secara optimal tanpa terlalu bergantung pada data pelatihan. Dengan demikian, algoritma Gradient Boosting terbukti efektif dalam mengenali hubungan antara variabel prediktor dengan tingkat risiko obesitas pada dataset medis yang digunakan.

Keberhasilan model dalam menghasilkan performa klasifikasi yang tinggi dipengaruhi oleh tahapan preprocessing yang dilakukan sebelum proses pemodelan. Tahap transformasi data kategorik menggunakan metode One-Hot Encoding memungkinkan data nonnumerik dapat diproses dengan baik oleh algoritma machine learning. Selain itu, proses normalisasi menggunakan Min-Max Scaler membantu menyamakan rentang nilai antar atribut sehingga model dapat bekerja lebih stabil dan optimal. Menurut Han, Kamber, dan Pei, kualitas data merupakan faktor penting dalam proses data mining karena sangat memengaruhi hasil klasifikasi yang dihasilkan model. Oleh karena itu, proses preprocessing yang tepat dapat meningkatkan kemampuan model dalam mempelajari pola data secara lebih efektif.

Selain tahapan preprocessing, proses seleksi variabel juga memberikan kontribusi terhadap peningkatan performa model. Variabel height dan weight dihapus dari proses klasifikasi karena kedua atribut tersebut digunakan secara langsung dalam perhitungan BMI untuk menentukan kategori obesitas. Jika kedua variabel tetap digunakan, maka model berpotensi mengalami data leakage, yaitu kondisi ketika model memperoleh informasi target secara tidak langsung sehingga menghasilkan akurasi yang terlalu tinggi namun tidak mencerminkan kemampuan model sebenarnya. Penghapusan atribut tersebut membuat model lebih fokus mempelajari pola dari variabel lain seperti kebiasaan makan, aktivitas fisik, konsumsi air, serta gaya hidup responden. Pendekatan ini menghasilkan model yang lebih realistis dan lebih representatif dalam menggambarkan kondisi sebenarnya.

Hasil eksplorasi variabel numerik menunjukkan bahwa sebagian besar variabel memiliki korelasi yang lemah. Hal tersebut mengindikasikan bahwa setiap atribut memberikan informasi yang berbeda terhadap proses klasifikasi. Dalam konteks

machine learning, rendahnya korelasi antarvariabel dapat membantu model mempelajari pola yang lebih kompleks karena tidak terjadi redundansi informasi. Korelasi yang paling menonjol ditemukan pada variabel height dan weight dengan hubungan positif sedang. Hal ini sesuai dengan teori bahwa berat badan cenderung meningkat seiring bertambahnya tinggi badan individu. Namun demikian, hubungan tersebut tidak cukup kuat untuk menjelaskan tingkat obesitas secara langsung tanpa mempertimbangkan faktor lain seperti pola makan, aktivitas fisik, serta kebiasaan hidup individu.

Berdasarkan hasil evaluasi Classification Report, nilai precision pada kelas tidak obesitas mencapai sekitar 93%, sedangkan pada kelas obesitas mencapai sekitar 88%. Nilai tersebut menunjukkan bahwa model cukup baik dalam meminimalkan kesalahan prediksi positif. Dalam bidang kesehatan, nilai precision yang tinggi penting untuk mengurangi kemungkinan kesalahan diagnosis terhadap individu yang sebenarnya tidak mengalami obesitas. Sementara itu, nilai recall pada kelas obesitas mencapai sekitar 92%, yang menunjukkan bahwa model mampu mengenali sebagian besar individu yang benar-benar termasuk kategori obesitas. Tingginya nilai recall menjadi penting dalam sistem kesehatan karena dapat membantu mengurangi risiko kegagalan deteksi pada individu yang membutuhkan penanganan lebih lanjut.

Nilai F1-score pada kedua kelas berada di sekitar 0,90 yang menunjukkan keseimbangan yang baik antara precision dan recall. Menurut Arthana, F1-score digunakan sebagai metrik evaluasi untuk mengukur keseimbangan performa model klasifikasi, terutama ketika diperlukan keseimbangan antara kemampuan mendeteksi kelas positif dan meminimalkan kesalahan prediksi. Dengan nilai F1-score yang tinggi, model dalam penelitian ini dapat dikatakan memiliki performa yang stabil dan seimbang pada kedua kelas klasifikasi.

Hasil Confusion Matrix menunjukkan bahwa model berhasil memprediksi 303 data negatif secara benar (true negative) dan 270 data positif secara benar (true positive). Namun demikian, masih terdapat 37 data false positive dan 24 data false negative. Kesalahan klasifikasi tersebut dapat dipengaruhi oleh kemiripan karakteristik antara individu obesitas dan tidak obesitas, keterbatasan atribut yang digunakan, serta kemungkinan adanya pola data yang sulit dipelajari secara sempurna oleh model. Meskipun demikian, jumlah kesalahan prediksi relatif kecil dibandingkan jumlah total data uji, sehingga performa model secara keseluruhan masih dapat dikategorikan sangat baik.

Algoritma Gradient Boosting dipilih dalam penelitian ini karena memiliki kemampuan membangun model secara bertahap melalui kombinasi beberapa weak learner berbasis pohon keputusan. Setiap model baru dibentuk untuk memperbaiki kesalahan dari model sebelumnya sehingga menghasilkan prediksi yang lebih akurat. Menurut Friedman, metode boosting mampu meningkatkan performa klasifikasi melalui proses minimisasi loss function secara iteratif. Pendekatan tersebut membuat

Gradient Boosting lebih efektif dibandingkan metode klasifikasi tunggal karena mampu mempelajari hubungan nonlinear dan pola kompleks antar variabel.

Hasil penelitian ini juga sejalan dengan penelitian sebelumnya yang dilakukan oleh Airlangga yang menunjukkan bahwa metode ensemble learning memiliki performa klasifikasi yang tinggi dalam prediksi obesitas berbasis data survei kesehatan. Penelitian lain oleh Syahputra dkk. juga menunjukkan bahwa algoritma machine learning seperti Random Forest, SVM, dan Gradient Boosting mampu memberikan hasil klasifikasi yang baik pada data kesehatan dengan tingkat akurasi yang tinggi. Kesamaan hasil tersebut memperkuat bahwa metode ensemble learning merupakan pendekatan yang efektif dalam menyelesaikan permasalahan klasifikasi pada bidang kesehatan.

Selain memberikan performa yang baik, penggunaan machine learning dalam klasifikasi risiko obesitas memiliki potensi implementasi yang luas dalam sistem informasi kesehatan. Model yang dibangun dapat digunakan sebagai alat bantu prediksi dini terhadap individu yang memiliki risiko obesitas berdasarkan pola aktivitas dan gaya hidup. Dengan adanya sistem prediksi berbasis data, tenaga medis dapat melakukan tindakan pencegahan lebih awal melalui edukasi pola makan sehat, peningkatan aktivitas fisik, serta pemantauan kesehatan secara berkala. Pemanfaatan machine learning dalam bidang kesehatan juga dapat membantu proses pengambilan keputusan menjadi lebih cepat, objektif, dan berbasis data historis.

Penelitian ini memiliki kontribusi dalam penerapan klasifikasi biner risiko obesitas menggunakan algoritma Gradient Boosting dengan pendekatan preprocessing untuk menghindari data leakage pada atribut berbasis BMI. Pendekatan tersebut membantu model menghasilkan performa klasifikasi yang lebih realistis dan stabil. Selain itu, penelitian ini menunjukkan bahwa kombinasi transformasi data, normalisasi, dan metode ensemble learning mampu meningkatkan efektivitas klasifikasi pada data medis yang memiliki karakteristik kompleks.

Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan berasal dari data publik dengan jumlah atribut yang terbatas sehingga belum sepenuhnya merepresentasikan kondisi kesehatan masyarakat secara menyeluruh. Selain itu, penelitian hanya menggunakan satu algoritma klasifikasi tanpa melakukan perbandingan langsung dengan metode lain seperti Random Forest, XGBoost, maupun Support Vector Machine. Oleh karena itu, penelitian selanjutnya dapat melakukan perbandingan beberapa algoritma machine learning untuk memperoleh metode terbaik dalam klasifikasi risiko obesitas. Penambahan jumlah dataset, penggunaan teknik feature selection, serta optimasi parameter model juga dapat dilakukan untuk meningkatkan performa klasifikasi di masa mendatang.

Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma Gradient Boosting mampu menjadi metode yang efektif dalam mengklasifikasikan risiko obesitas berdasarkan data medis. Kombinasi antara preprocessing data, transformasi variabel, pemetaan kelas, serta penggunaan metode ensemble learning menghasilkan model

dengan performa tinggi dan stabil. Hasil penelitian ini diharapkan dapat menjadi kontribusi dalam pengembangan sistem prediksi kesehatan berbasis kecerdasan buatan yang lebih akurat, adaptif, dan bermanfaat bagi dunia medis maupun penelitian selanjutnya.

KESIMPULAN

Berdasarkan penelitian, variabel respons pada dataset obesity yang awalnya terdiri atas tujuh kelas disederhanakan menjadi dua kelas utama, yaitu kelas 0 (tidak obesitas) dan kelas 1 (obesitas), untuk memfokuskan klasifikasi biner. Sebelum pemodelan, data melalui pre-processing meliputi analisis korelasi antarvariabel, transformasi dan normalisasi variabel, pemetaan kelas, serta pemisahan ke variabel independen (X) dan dependen (y) untuk memastikan model dapat mempelajari pola secara optimal. Pemrosesan data menggunakan Gradient Boosting Classifier dengan 100 estimator, learning rate 0,1, dan kedalaman pohon maksimum 4 menghasilkan akurasi 90,38% tanpa indikasi overfitting atau underfitting, serta precision, recall, dan f1-score yang seimbang pada kedua kelas. Hasil ini juga menekankan pentingnya pemilihan variabel yang relevan untuk meningkatkan performa model.

Hasil ini menunjukkan Gradient Boosting mampu mengklasifikasikan dataset obesity dengan baik berdasarkan variabel yang ada. Sebagai metode ensemble, algoritma ini efektif mempelajari pola kompleks dan memiliki potensi diterapkan di bidang kesehatan untuk membantu tenaga medis mengidentifikasi individu dengan risiko obesitas secara lebih dini dan akurat.

DAFTAR PUSTAKA

[1] Aditya, N. R. (2015). *Data mining*. UNIKOM. https://repository.unikom.ac.id/47451/1/Pertemuan%203%20-%20Materi%20%5BDM%20-%202015%5D.pdf

[2] Airlangga, G. (2025). Machine learning-based obesity classification: A comparative study using self-reported survey data and ensemble learning models. *Jurnal Teknologi Informatika dan Komputer, 11*(1).

[3] Arrahimi, A. R., Ihsan, M. K., Kartini, D., Faisal, M. R., & Indriani, F. (2019). Teknik bagging dan boosting pada algoritma CART untuk klasifikasi masa studi mahasiswa. *Jurnal Sains dan Informatika, 5*. https://jsi.politala.ac.id/index.php/JSI/article/view/171/94

[4] Arthana, R. (2019, April 5). Mengenal accuracy, precision, recall dan specificity serta yang diprioritaskan. *Medium*. https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

[5] Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan akurasi, recall, dan presisi klasifikasi pada algoritma C4.5, random forest, SVM, dan naive bayes. *Jurnal Media Informatika Budidarma, 5*, 640–651. https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/2937

[6] Direktorat Pengendalian Penyakit Tidak Menular. (2015). *Pedoman umum pengendalian obesitas*. Kementerian Kesehatan Republik Indonesia. https://extranet.who.int/ncdccs/Data/IDN_B11_Buku%20Obesitas-1.pdf

[7] EKRUT. (2022, September 28). Dataset adalah: Pengertian, tipe, perbedaan dengan database, dan 10 web penyedia. https://www.ekrut.com/media/dataset-adalah

[8] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

[9] Mendoza, F. P., & de la H. Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief,* 25, 104344. https://doi.org/10.1016/j.dib.2019.104344

[10] Pujiastuti, P. (2012). Obesitas dan penyakit periodontal. *Stomatognatic (J.K.G Unej),* 9, 82–85. https://jurnal.unej.ac.id/index.php/STOMA/article/download/2112/1715

[11] Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan machine learning dalam berbagai bidang: Review paper. *IJCIT,* 5, 75–82. https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit/article/view/7951/pdf

[12] Shalini, K., Shanthi, A. V. K., Shakila, C., & Chamudeeswari, N. (2025). Machine learning approaches for obesity level classification. *International Journal of Environmental Sciences, 11.*

[13] Syahputra, A. R., Hidayat, R., Rismansyah, F., et al. (2025). Komparasi algoritma machine learning (SVM, random forest, dan regresi logistik) untuk prediksi tingkat obesitas. *Jurnal Ilmiah Teknik Informatika dan Komunikasi,* 5(3).

[14] Tibshirani, R., Friedman, J., & Hastie, T. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[15] Yulianti, I. F., & Sihombing, P. R. (2021). Penerapan metode machine learning dalam klasifikasi risiko kejadian berat badan lahir rendah di Indonesia. *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer,* 20, 417–426. https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/1174/703