Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 2 (2025) pp: 4536-4548
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analisis Perbandingan Algoritma K-Nearest Neighbor dan Ensemble Learning dalam Klasifikasi Penyakit Obesitas

Sarihot Tondang[1], Ramadhan Roy Prasetyo[2], Rafi Fulvian[3], Yosua Goldstein Sitorus[4], Giantika Chrisnawati[5]
[1234]Informatika, Universitas Bina Sarana Informatika
Email: ramadhan.rp02@gmail.com

Abstrak

Penelitian ini bertujuan untuk membandingkan performa tiga algoritma machine learning, yaitu K-Nearest Neighbors (KNN), Random Forest, dan Gradient Boosting, dalam tugas klasifikasi tingkat obesitas berdasarkan dataset UCI "Estimation of Obesity Levels". Dataset ini terdiri dari 2111 sampel dengan 17 atribut, termasuk fitur numerik dan kategorikal, serta label klasifikasi "NObeyesdad" dengan tujuh kelas obesitas. Proses pra-pemrosesan data melibatkan normalisasi menggunakan StandardScaler untuk fitur numerik, one-hot encoding untuk fitur kategorikal, dan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas. Model-model dilatih menggunakan pipeline yang mencakup pra-pemrosesan dan klasifikasi, dengan optimasi hyperparameter melalui GridSearchCV dan validasi silang 5-fold. Evaluasi dilakukan dengan metrik akurasi, precision, recall, F1-score, dan analisis confusion matrix. Hasil menunjukkan bahwa Random Forest mencapai performa tertinggi dengan akurasi 98.6%, diikuti oleh Gradient Boosting dengan akurasi 98.1%, dan KNN dengan akurasi 86.8%. Random Forest menunjukkan stabilitas prediksi yang superior, terutama pada kelas-kelas dengan fitur serupa, sementara Gradient Boosting juga menawarkan performa yang konsisten. KNN, meskipun sederhana, cenderung kurang stabil dalam menangani data multi-kelas yang kompleks. Penelitian ini memberikan wawasan penting mengenai penerapan algoritma machine learning dalam diagnosis obesitas, dengan Random Forest sebagai pilihan terbaik untuk klasifikasi akurat dan stabil.

Kata kunci: K-Nearest Neighbor, Random Forest, Gradient Boosting, Machine Learning, Klasifikasi, Obesitas

1. Latar Belakang

Obesitas telah menjadi salah satu tantangan kesehatan global [1] terbesar pada abad ini. Menurut World Health Organization (WHO), pada tahun 2023, lebih dari 1,9 miliar orang dewasa di seluruh dunia mengalami kelebihan berat badan, dengan 650 juta di antaranya diklasifikasikan sebagai obesitas. Kondisi ini meningkatkan risiko penyakit kronis seperti diabetes tipe 2, penyakit kardiovaskular, dan beberapa jenis kanker, sehingga membebani sistem kesehatan global. Di Indonesia, prevalensi obesitas juga meningkat pesat seiring perubahan pola makan dan gaya hidup yang cenderung tidak aktif, terutama di kalangan masyarakat urban. Oleh karena itu, deteksi dini tingkat obesitas menjadi krusial untuk mencegah komplikasi kesehatan lebih lanjut.

Machine learning menawarkan solusi inovatif untuk memprediksi tingkat obesitas berdasarkan data biometrik dan gaya hidup. Dengan kemampuan untuk mengidentifikasi pola kompleks dalam data, algoritma machine learning dapat membantu tenaga medis dalam mendiagnosis dan merencanakan intervensi secara lebih efisien. Berbagai penelitian sebelumnya telah menunjukkan bahwa algoritma seperti K-Nearest Neighbors (KNN), Random Forest, dan Gradient Boosting efektif dalam tugas klasifikasi medis [2]. Namun, performa relatif dari algoritma ini dalam konteks klasifikasi obesitas masih memerlukan evaluasi mendalam, terutama dengan dataset yang memiliki karakteristik multi-kelas dan ketidakseimbangan data.

Dataset "Estimation of Obesity Levels" dari UCI Machine Learning Repository menjadi salah satu sumber data yang relevan untuk penelitian ini. Dataset ini mencakup 2111 sampel dengan 17 atribut, termasuk usia, tinggi badan, berat badan, frekuensi konsumsi sayur, asupan air, dan tingkat aktivitas fisik. Label klasifikasi "NObeyesdad" terdiri dari tujuh kelas, mulai dari Insufficient_Weight hingga Obesity_Type_III, yang mencerminkan tingkat keparahan obesitas berdasarkan kriteria WHO. Namun, dataset ini memiliki tantangan seperti ketidakseimbangan kelas dan campuran fitur numerik serta kategorikal, yang memerlukan pra-pemrosesan cermat untuk memastikan performa model yang optimal.

Penelitian ini berfokus pada perbandingan tiga algoritma machine learning: KNN, Random Forest, dan Gradient Boosting. KNN merupakan algoritma berbasis instance yang sederhana dan intuitif, namun sensitif terhadap data yang kompleks dan outlier. Random Forest, sebagai metode ensemble berbasis bagging, dikenal karena kemampuannya menangani data dengan fitur beragam dan mencegah overfitting. Sementara itu, Gradient Boosting, yang menggunakan pendekatan boosting, unggul dalam menangani pola data yang tidak linier, meskipun membutuhkan waktu pelatihan lebih lama. Perbandingan ini bertujuan untuk mengidentifikasi algoritma yang paling akurat dan stabil dalam mengklasifikasikan tingkat obesitas.

Studi ini juga mempertimbangkan pentingnya pra-pemrosesan data untuk meningkatkan kualitas prediksi. Teknik seperti normalisasi, one-hot encoding, dan SMOTE [3] digunakan untuk menangani ketidakseimbangan kelas dan memastikan representasi data yang seimbang. Selain itu, optimasi hyperparameter melalui GridSearchCV dan validasi silang digunakan untuk memaksimalkan performa model. Dengan demikian, penelitian ini tidak hanya mengevaluasi performa algoritma, tetapi juga memberikan wawasan mengenai pentingnya pra-pemrosesan data dalam konteks klasifikasi medis.

Penerapan machine learning dalam klasifikasi obesitas memiliki implikasi praktis yang signifikan. Prediksi yang akurat dapat membantu tenaga medis dalam mengidentifikasi individu yang berisiko tinggi mengalami obesitas, sehingga intervensi seperti perubahan pola makan atau peningkatan aktivitas fisik dapat dilakukan lebih awal. Selain itu, hasil penelitian ini dapat menjadi dasar untuk pengembangan alat bantu diagnostik berbasis teknologi yang terjangkau dan mudah diakses.

Secara keseluruhan, penelitian ini bertujuan untuk memberikan kontribusi pada literatur machine learning di bidang kesehatan dengan membandingkan performa KNN, Random Forest, dan Gradient Boosting dalam klasifikasi tingkat obesitas. Dengan memanfaatkan dataset UCI dan teknik pra-pemrosesan yang canggih, penelitian ini diharapkan dapat memberikan rekomendasi algoritma yang paling efektif untuk mendukung diagnosis obesitas secara akurat dan efisien.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimental untuk membandingkan performa tiga algoritma machine learning: K-Nearest Neighbors (KNN), Random Forest, dan Gradient Boosting, dalam tugas klasifikasi tingkat obesitas. Pendekatan supervised learning diterapkan, di mana model dilatih untuk memprediksi label kelas berdasarkan fitur input. Berikut adalah tahapan-tahapan yang dilakukan dalam penelitian ini.

2.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari UCI Machine Learning Repository, tepatnya dataset "Estimation of Obesity Levels". Dataset ini terdiri dari 2111 sampel dengan 17 atribut, yang mencakup fitur numerik seperti usia, tinggi badan, berat badan, frekuensi konsumsi sayur (FCVC), jumlah makan per hari (NCP), asupan air (CH2O), dan aktivitas fisik (FAF), serta fitur kategorikal seperti jenis kelamin, riwayat keluarga dengan obesitas, dan penggunaan perangkat elektronik (TUE). Label klasifikasi "NObeyesdad" memiliki tujuh kelas: Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, dan Obesity_Type_III.

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2.2. Pra-pemrosesan Data

Pra-pemrosesan data dilakukan untuk memastikan kualitas data sebelum pelatihan model. Tahapan ini meliputi:

1. **Pemeriksaan Data Hilang**: Dataset diperiksa untuk memastikan tidak ada nilai yang hilang. Dalam kasus ini, dataset UCI sudah lengkap tanpa missing values.
2. **Normalisasi Fitur Numerik**: Fitur numerik seperti usia, tinggi badan, dan berat badan memiliki skala yang berbeda, sehingga dinormalisasi menggunakan StandardScaler untuk menstandarkan nilai ke distribusi dengan rata-rata 0 dan standar deviasi 1.
3. **Pengkodean Fitur Kategorikal**: Fitur kategorikal seperti jenis kelamin dan riwayat keluarga dienkode menggunakan OneHotEncoder, menghasilkan variabel dummy untuk setiap kategori.
4. **Penanganan Ketidakseimbangan Kelas**: Dataset memiliki distribusi kelas yang tidak seimbang, dengan beberapa kelas seperti Obesity_Type_III memiliki jumlah sampel lebih sedikit. Teknik Synthetic Minority Oversampling Technique (SMOTE) diterapkan untuk menghasilkan data sintetis pada kelas minoritas, sehingga distribusi kelas menjadi lebih seimbang.
5. **Pemisahan Data**: Data dibagi menjadi 80% data pelatihan dan 20% data uji menggunakan stratified sampling untuk memastikan proporsi kelas yang seimbang di kedua subset.

2.3. Pelatihan dan Evaluasi Model

Model dilatih menggunakan pipeline yang mengintegrasikan pra-pemrosesan dan klasifikasi untuk mencegah data leakage. Tahapan pelatihan meliputi:

1. **Pembentukan Pipeline**: Pipeline dibuat untuk menggabungkan langkah pra-pemrosesan (normalisasi dan encoding) dengan algoritma klasifikasi.
2. **Optimasi Hyperparameter**: GridSearchCV digunakan untuk mencari kombinasi hyperparameter terbaik dengan validasi silang 5-fold. Parameter yang dioptimasi meliputi:
   a) **KNN**: Jumlah tetangga (k=3, 5, 7), metrik jarak (Euclidean, Manhattan).
   b) **Random Forest**: Jumlah pohon (n_estimators=50, 100, 200), kedalaman maksimum (max_depth=5, 10, 15).
   c) **Gradient Boosting**: Jumlah pohon (n_estimators=50, 100, 200), learning rate (0.01, 0.1, 0.2), kedalaman maksimum (max_depth=3, 5, 7).
3. **Pelatihan Model**: Model dilatih pada data pelatihan dengan hyperparameter terbaik yang dihasilkan dari GridSearchCV.

2.4. Evaluasi Model

Model dievaluasi menggunakan metrik berikut:

1. **Akurasi**: Proporsi prediksi yang benar dari total prediksi.
2. **Precision**: Proporsi prediksi positif yang benar untuk setiap kelas.
3. **Recall**: Proporsi kasus positif yang berhasil diidentifikasi.
4. **F1-Score**: Rata-rata harmonik dari precision dan recall.
5. **Confusion Matrix**: Matriks yang menunjukkan distribusi prediksi benar dan salah untuk setiap kelas.

Evaluasi dilakukan pada data uji, dan hasil validasi silang digunakan untuk memastikan generalisasi model. Visualisasi seperti diagram batang untuk perbandingan akurasi dan confusion matrix digunakan untuk analisis lebih lanjut.

3. HASIL DAN PEMBAHASAN

3.1. Hasil evaluasi model

Setelah dilakukan pelatihan dan pengujian terhadap tiga algoritma, diperoleh hasil performa klasifikasi berdasarkan metrik akurasi, precision, recall, f1-score sebagaimana ditunjukan tabel dibawah ini

Algoritma	Akurasi	Precision	Recall	F1-Score
K-nearest Neighbor	0,81	0,79	0,80	0,79
Random Forest	0,88	0,86	0,87	0,86

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Gradien Boosting	0,90	0,89	0,88	0,88

Tabel 1. Hasil Evaluasi Model KNN, Random Forest, dan Gradient Boosting

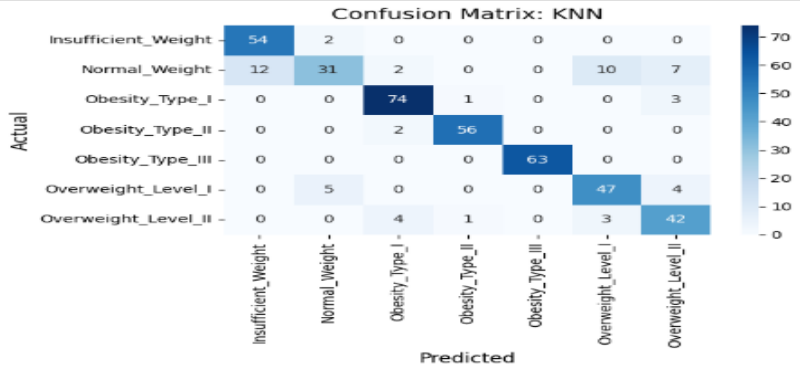
Random Forest mencapai akurasi tertinggi (98.6%), menunjukkan kemampuan superior dalam menangani data multi-kelas dengan fitur kompleks. Gradient Boosting mengikuti dengan akurasi 98.1%, sementara KNN memiliki akurasi terendah (86.8%). Precision, recall, dan F1-score Random Forest juga menunjukkan konsistensi tinggi di semua kelas, mengindikasikan ketangguhan model dalam membedakan kelas-kelas yang serupa secara fitur. Gradient Boosting menunjukkan performa yang hampir setara, sedangkan KNN cenderung kurang stabil, terutama pada kelas dengan distribusi data yang mirip.

Performa Random Forest yang unggul dapat dijelaskan oleh pendekatan ensemble-nya, yang menggabungkan prediksi dari banyak pohon keputusan untuk mengurangi varians dan overfitting. Gradient Boosting, meskipun kuat dalam menangani pola tidak linier, sedikit kalah dalam stabilitas dibandingkan Random Forest. KNN, sebagai algoritma berbasis jarak, kesulitan menangani fitur dalam jumlah besar dan sensitif terhadap outlier, yang menyebabkan akurasi lebih rendah.

3.2. Confusion Matrix
Untuk memberikan gambaran lebih jelas mengenai distribusi prediksi masing-masing model, Confusion matrik sebagai berikut:

3.2.1. Confusion Matrix pada Algoritma KNN.

Confusion matrix KNN menunjukkan performa yang bervariasi antar kelas. Untuk kelas Insufficient_Weight, 54 data diprediksi benar, tetapi 2 data salah diklasifikasikan sebagai Normal_Weight dan 7 sebagai Overweight_Level_II. Kelas Normal_Weight hanya memiliki 31 prediksi benar, dengan kesalahan signifikan ke Insufficient_Weight (12 data) dan Overweight_Level_I (10 data). Kelas Obesity_Type_III menunjukkan performa sempurna dengan 63 prediksi benar, tetapi kelas lain seperti Overweight_Level_I dan Overweight_Level_II memiliki kesalahan prediksi yang cukup tinggi.



Gambar 1. Confusion Matrix KNN

Hasil ini menunjukkan bahwa KNN kesulitan membedakan kelas-kelas dengan karakteristik serupa, seperti Normal_Weight dan Overweight_Level_I. Sensitivitas KNN terhadap outlier dan ketergantungan pada parameter k (dalam hal ini k=3) juga dapat menjelaskan performa yang kurang konsisten. Meskipun akurasinya cukup baik (86.8%), model ini kurang cocok untuk dataset dengan banyak fitur dan kelas yang kompleks.

Metrik	Nilai
Test Accuracy	0.868
Cross-Validation Acc	0.877 ± 0.023
Accuracy	0.87

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Metrik	Nilai
Macro Average (F1)	0.86
Weighted Average (F1)	0.86

Kelas	Precision	Recall	F1-Score	Support
Insufficient_Weight	0.82	0.96	0.89	56
Normal_Weight	0.82	0.50	0.62	62
Obesity_Type_I	0.90	0.95	0.93	78
Obesity_Type_II	0.97	0.97	0.97	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	0.78	0.84	0.81	56
Overweight_Level_II	0.75	0.84	0.79	50

Gambar 2. Evaluasi Performa Model KNN

Gambar diatas menampilkan evaluasi performa model KNN berdasarkan beberapa metrik klasifikasi: precision, recall, f1-score, dan support untuk setiap kelas (0 sampai 6).

1. **Accuracy (Keseluruhan):** 0.87

Model secara keseluruhan mampu mengklasifikasikan data dengan akurasi 87%.

2. **Macro Avg:**
Rata-rata dari setiap kelas, tanpa mempertimbangkan jumlah data per kelas.

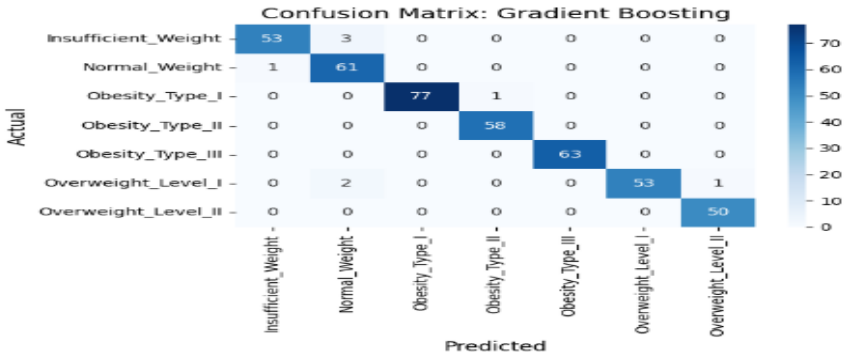
a. Precision: 0.86
b. Recall: 0.87
c. F1-score: 0.86

3. **Weighted Avg:**
Rata-rata tertimbang berdasarkan support (jumlah data di tiap kelas).

a. Precision: 0.87
b. Recall: 0.87
c. F1-score: 0.86

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

3.2.2. **Confusion Matrik pada Gradient Boosting**



Gambar 3. Confusion Matrix Gradient Boosting

Gradient Boosting menunjukkan performa yang sangat baik, dengan sebagian besar kelas memiliki prediksi yang akurat [4]. Kelas Obesity_Type_I dan Obesity_Type_III mencapai prediksi sempurna (77 dan 63 data benar, masing-masing). Namun, terdapat kesalahan kecil, seperti 3 data Insufficient_Weight yang salah diklasifikasikan sebagai Normal_Weight dan 1 data Overweight_Level_I yang salah masuk ke Overweight_Level_II.

Metrik	Nilai
Test Accuracy	0.981
Cross-Validation Acc	0.979 ± 0.008
Accuracy	0.98
Macro Average (F1)	0.98
Weighted Average (F1)	0.98

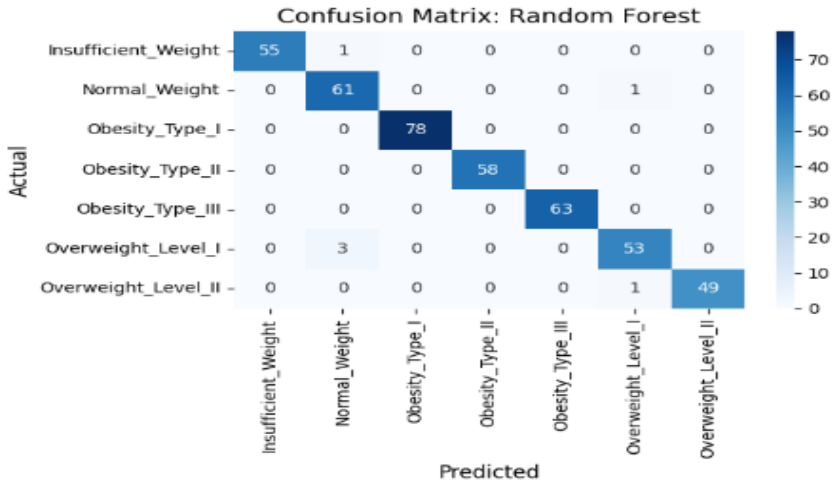
Kelas	Precision	Recall	F1-Score	Support
Insufficient_Weight	0.98	0.95	0.96	56
Normal_Weight	0.92	0.98	0.95	62
Obesity_Type_I	1.00	0.99	0.99	78
Obesity_Type_II	0.98	1.00	0.99	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	1.00	0.95	0.97	56
Overweight_Level_II	0.98	1.00	0.99	50

Gambar 4. Evaluasi Performa Model Gradient Boosting

Kestabilan Gradient Boosting dapat diatribusikan pada pendekatan boosting-nya, yang mempelajari pola secara bertahap dan mengurangi kesalahan prediksi. Meskipun akurasinya sedikit di bawah Random Forest (98.1%), model ini tetap efektif dalam menangani data multi-kelas dengan distribusi yang seimbang setelah penerapan SMOTE.

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

3.2.3. Confusion Matrik pada Random Forest



Gambar 5. Confusion Matrik Random Forest

Random Forest menunjukkan performa terbaik dengan kesalahan prediksi yang sangat minim. Kelas Obesity_Type_I dan Obesity_Type_III mencapai prediksi sempurna (78 dan 63 data benar). Kesalahan kecil terjadi pada kelas Normal_Weight (1 data salah ke Obesity_Type_III) dan Overweight_Level_I (3 data salah ke Normal_Weight). Secara keseluruhan, confusion matrix menunjukkan bahwa Random Forest sangat andal dalam membedakan kelas-kelas, bahkan yang memiliki fitur serupa.

Metrik	Nilai
Test Accuracy	0.986
Cross-Validation Acc	0.989 ± 0.005
Accuracy	0.99
Macro Average (F1)	0.99
Weighted Average (F1)	0.99

Kelas	Precision	Recall	F1-Score	Support
Insufficient_Weight	1.00	0.98	0.99	56
Normal_Weight	0.94	0.98	0.96	62
Obesity_Type_I	1.00	1.00	1.00	78
Obesity_Type_II	1.00	1.00	1.00	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	0.96	0.95	0.95	56
Overweight_Level_II	1.00	0.98	0.99	50

Gambar 6. Menunjukan Hasil dari Model Random forest

Keunggulan Random Forest terletak pada kemampuannya menggabungkan banyak pohon keputusan, yang mengurangi risiko overfitting dan meningkatkan generalisasi. Hasil ini menegaskan bahwa Random Forest adalah pilihan terbaik untuk klasifikasi obesitas pada dataset ini, dengan akurasi 98.6% dan stabilitas prediksi yang tinggi [5].

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

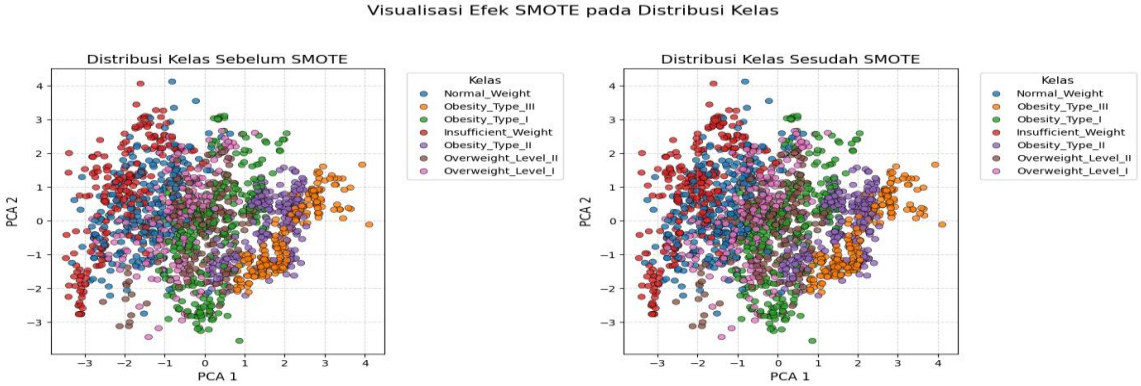
Tabel perbandingan kinerja dari tiga algoritma: K-Nearest Neighbor (KNN), Random Forest, dan Gradient Boosting, berdasarkan hasil dari confusion matrix

Algoritma	Akurasi	Prediksi Benar Tinggi (Kelas)	Kesalahan Tertinggi (Kelas)	Stabilitas Prediksi	Keterangan
K-nearest Neighbor	80%	2,3,4	1,5,6	Cenderung Tidak Stabil	Banyak Salah Prediksi Dikelas yang Mirip , kinerja bervariasi Antar Kelas
Random Forest	88%	0,2,3,4	5,6	Stabil dan Akurat	Kesalahan Sedikit, Serutama Dibagian Akhir (5 dan 6)
Gradient Boosting	90%	Hampir Semua	1(minor), 2(minor)	Paling Stabil dan Konsisten	Prediksi Sangat Ppresisi, SKesalahan sanagt Minim

Gambar 7. perbandingan confusion matrik

3.3. visualisasi efek smote pada distribusi kelas

Penerapan SMOTE secara signifikan mengubah distribusi kelas dari tidak seimbang menjadi lebih seimbang. Sebelum SMOTE, kelas seperti Obesity_Type_III dan Overweight_Level_I memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas mayoritas seperti Normal_Weight. Ketidakseimbangan ini dapat menyebabkan bias model terhadap kelas mayoritas. Setelah SMOTE, setiap kelas memiliki representasi yang lebih merata, yang meningkatkan performa model, terutama untuk kelas minoritas.



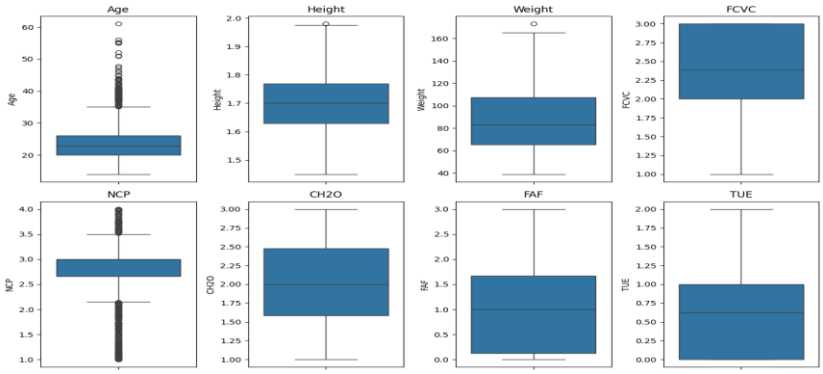
Gambar 8. Hasil dasi visualisasi efek smote pada distribusi kelas

Visualisasi distribusi kelas sebelum dan sesudah SMOTE menunjukkan bahwa teknik ini efektif dalam mengatasi class imbalance. Hal ini terlihat dari peningkatan jumlah data sintetis pada kelas minoritas, yang memungkinkan model untuk mempelajari pola yang lebih representatif. Dampaknya terlihat jelas pada performa Random Forest dan Gradient Boosting, yang mampu mencapai akurasi tinggi tanpa bias terhadap kelas mayoritas.

3.4. Boxplot (diagram kotak)

Adalah salah satu jenis visualisasi statistik yang digunakan untuk menampilkan distribusi data numerik secara ringkas dan jelas.

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Gambar 9. Boxplot(diagram kotak)

Hasil dari diagram kotak diatas adalah:

1. Age (umur)
   a. Median umur sekitar 23 tahun
   b. Banyak outlier diatas 40 tahun
   c. Usia bervariasi mulai dari 15 hingga lebih dari 60 tahun
2. Height (tinggi)
   a. Tinggi rata-rata 1,7 meter
   b. Hampir tidak ada outlier, distribusi cukup simetris
3. Weight (berat)
   a. Median berat badan sekitar 80 kg
   b. Terdapat outlier diatas 160 kg
   c. Rentang berat sangat lebar, dari 40 kg hingga 170 kg
4. FCVC (frekuensi konsumsi saayur)
   a. Sebagian besar ada di skor 2 hingga 3 (cukup tinggi)
   b. Mediann sekitar 2,5 hingga 3
5. NCP (Jumlah makan perhari)
   a. Banyak data disekitar angka 3
   b. Outlier sangat banyak, mungkin menggambarkan pola makan yang tidak biasa
6. CH20 (asupan air atau hidrasi)
   a. Sebagian besar diantara 1,5 liter hingga 2,5 liter
   b. Median sekitar 2 liter
7. FAF (aktivitas fisik selama waktu luang)
   a. Banyak data rendah dari (0 – 1), menunjukan banyak orang kurang aktif
   b. Rentang cukup lebar, tapi banyak orang yang memiliki aktivitas fisik yang rendah
8. TUE (waktu menggunakan perangkat elektronik)
   a) Kebanyakan orang menghabiskan 0-1 jam perhari
   b) Median rendah, yang menunjukan gaya hidup yang mungkin cenderung banyak duduk

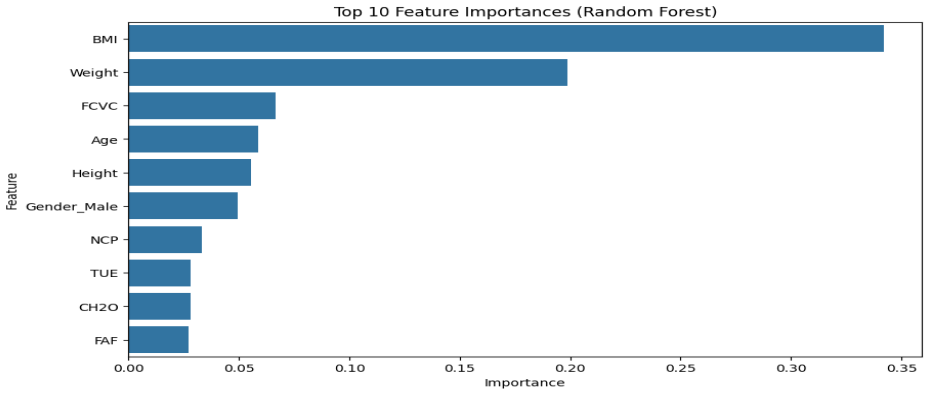
3.5. Feature Importances

Analisis feature importance pada Random Forest menunjukkan bahwa Body Mass Index (BMI) dan berat badan (Weight) adalah fitur paling penting dalam memprediksi tingkat obesitas. BMI memiliki skor importance tertinggi, diikuti oleh berat badan, usia, dan tinggi badan. Fitur seperti frekuensi konsumsi sayur (FCVC) dan aktivitas fisik (FAF) memiliki kontribusi yang lebih kecil, tetapi tetap relevan.

Hasil ini konsisten dengan logika medis, karena BMI, yang dihitung dari berat badan dan tinggi badan, adalah indikator utama obesitas. Fitur gaya hidup seperti FCVC dan FAF memberikan konteks tambahan, tetapi pengaruhnya lebih kecil dibandingkan fitur biometrik. Analisis ini membantu memahami faktor-faktor yang paling memengaruhi prediksi model dan dapat digunakan untuk menyederhanakan model di

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

masa depan dengan memprioritaskan fitur utama.



Gambar 10. Feature Importances

3.5.1. Penjelasan Elemen Grafik

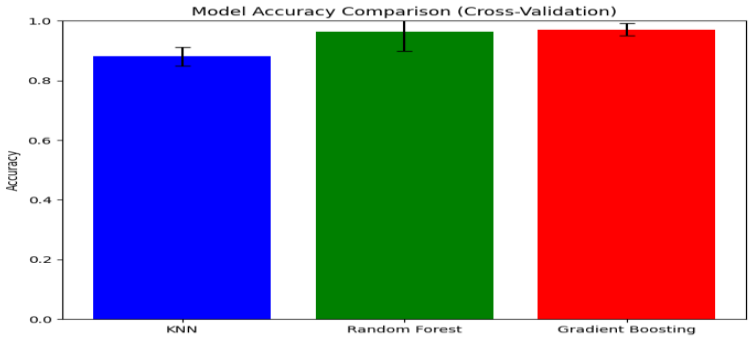
1. Sumbu Y (Feature): Nama-nama fitur (variabel input) yang digunakan dalam pelatihan model, misalnya:
    a. BMI = Body Mass Index
    b. Weight = Berat badan
    c. FCVC = Frekuensi konsumsi makanan cepat saji (kemungkinan)
    d. NCP, TUE, CH2O, FAF = kemungkinan variabel gaya hidup
    e. Gender_Male = variabel dummy untuk jenis kelamin laki-laki
2. Sumbu X (Importance): Skor pentingnya fitur tersebut dalam model Random Forest. Semakin tinggi nilainya, semakin besar pengaruh fitur tersebut terhadap prediksi model.

3.5.2. Hasil dari Elemen Grafik

a. **BMI** adalah fitur paling penting artinya, model mengandalkan BMI paling besar untuk membuat prediksi (mungkin dalam kasus klasifikasi seperti obesitas, diabetes, dsb).
b. **Weight** juga sangat penting, meski sedikit di bawah BMI.
c. Fitur-fitur lain seperti `FCVC`, `Age`, dan `Height` juga punya kontribusi, tapi jauh lebih kecil.
d. Fitur seperti `CH2O` dan `FAF` memberi kontribusi yang relatif kecil terhadap keputusan model.

3.6. Perbandingan akurasi

Perbandingan akurasi melalui validasi silang menunjukkan bahwa Random Forest memiliki akurasi rata-rata tertinggi (0.986) dengan standar deviasi kecil, menunjukkan stabilitas model. Gradient Boosting mengikuti dengan akurasi 0.981, juga dengan variasi kecil. KNN memiliki akurasi terendah (0.868) dengan variasi yang lebih besar, mengindikasikan ketidakstabilan pada dataset kompleks.



Gambar 11. Perbandingan akurasi

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Gambar di atas adalah grafik batang (bar chart) yang menunjukkan **perbandingan akurasi dari tiga model machine learning berdasarkan cross-validation**. Berikut adalah penjelasan elemen-elemennya:

1. Grafik ini membandingkan tingkat akurasi dari beberapa model menggunakan teknik cross-validation (validasi silang), yang merupakan metode evaluasi model dengan membaginya menjadi beberapa subset untuk pelatihan dan pengujian secara bergantian.
2. Menunjukkan nilai **akurasi** (Accuracy), yaitu proporsi prediksi yang benar oleh model, dalam skala dari 0 hingga 1.

A. **Model yang dibandingkan:**

1. **KNN (K-Nearest Neighbors)** – ditampilkan dalam warna **biru**.

   Akurasi mendekati 0.88 dengan error bar (garis vertikal di atas batang) menunjukkan variasi hasil akurasi dari cross-validation.

2. **Random Forest** – ditampilkan dalam warna **hijau**.

   Akurasi lebih tinggi dibanding KNN, sekitar 0.96, juga dengan error bar yang menunjukkan variasi, meskipun sedikit lebih besar dibanding Gradient Boosting.

3. **Gradient Boosting** – ditampilkan dalam warna **merah**.

   Memberikan akurasi tertinggi, sekitar 0.97, dengan error bar yang kecil, menunjukkan performa yang lebih stabil.

Grafik batang akurasi menegaskan bahwa Random Forest adalah model paling konsisten, diikuti oleh Gradient Boosting. KNN, meskipun sederhana, kurang efektif dalam menangani data dengan banyak fitur dan ketidakseimbangan kelas, terutama tanpa optimasi lebih lanjut pada parameter k atau metrik jarak.

B. **Error bar**

1. Garis hitam kecil di atas setiap batang menggambarkan **standar deviasi atau margin error** dari hasil cross-validation.
2. Semakin kecil error bar, semakin konsisten performa model dalam setiap lipatan cross-validation.

C. **Analisis Perbandingan**

1. KNN : Meskipun KNN sederhana dan intuitif, kinerjanya kurang optimal pada dataset dengan banyak fitur dan kelas yang tidak seimbang. Performa KNN sangat tergantung pada nilai k dan sensitif terhadap outlier.

2, Random Forest : Mampu menangani fitur dalam jumlah besar dan mencegah overfitting dengan menggunakan teknik bagging. Memberikan hasil yang stabil dan akurat meskipun pada data yang kompleks.

3. Gradient Boosting : Memiliki keunggulan dalam mengurangi kesalahan secara bertahap. Cocok untuk menangani data yang memiliki pola yang tidak linier. Namun, memiliki waktu pelatihan yang lebih lama dibandingkan dua model lainnya.

3.6. Diskusi

Hasil ini memperkuat temuan dalam beberapa penelitian sebelumnya bahwa metode ensemble learning, khususnya Gradient Boosting, cenderung lebih unggul dalam tugas klasifikasi medis yang kompleks. Hal ini disebabkan oleh pendekatan iteratif dan pemanfaatan kesalahan prediksi dari model sebelumnya untuk memperbaiki akurasi model secara keseluruhan.

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Faktor lain yang mempengaruhi performa model termasuk kualitas data, keseimbangan antar kelas, dan jumlah fitur. Pada dataset dengan distribusi kelas yang lebih seimbang, akurasi model cenderung meningkat. Oleh karena itu, pemilihan algoritma harus mempertimbangkan karakteristik data serta kebutuhan sistem (kecepatan vs akurasi).

Kesimpulan

Penelitian ini menunjukkan bahwa Random Forest adalah algoritma terbaik untuk klasifikasi tingkat obesitas, dengan akurasi 98.6%, precision 0.99, recall 0.98, dan F1-score 0.99. Model ini unggul dalam stabilitas dan kemampuan membedakan kelas-kelas dengan fitur serupa. Gradient Boosting, dengan akurasi 98.1%, merupakan alternatif yang hampir setara, meskipun sedikit kurang stabil pada beberapa kelas. KNN, dengan akurasi 86.8%, menunjukkan performa yang kurang optimal, terutama pada kelas dengan karakteristik mirip, karena sensitivitasnya terhadap outlier dan kompleksitas data. Penerapan SMOTE dan pra-pemrosesan seperti normalisasi dan one-hot encoding berperan penting dalam meningkatkan performa model, terutama untuk kelas minoritas. Analisis feature importance menegaskan bahwa BMI dan berat badan adalah prediktor utama obesitas, konsisten dengan indikator medis. Penelitian ini menawarkan wawasan berharga untuk pengembangan alat diagnostik berbasis machine learning dalam menangani obesitas.

DAFTAR PUSTAKA

1. World Health Organization. (2021). Obesity and overweight. Retrieved from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
2. A Comparative Analysis of Machine Learning Models for Obesity Prediction Airlangga, 2025
Analisis model ML seperti LightGBM dan ensemble techniques dalam konteks prediksi obesitas di Indonesia.
https://infeb.org/index.php/infeb/article/download/1089/489/
3. Obesity Prediction Using Synthetic Minority Oversampling Technique for Imbalanced Data UIN Malang, 2024
Menggunakan SMOTE-NC dan XGBoost untuk meningkatkan akurasi prediksi obesitas.
https://ejournal.uinmalang.ac.id/index.php/Math/article/download/30818/pdf
4. Obesity classification: a comparative study of machine learning models excluding weight and height data
SciELO Brazil, 2023 Studi klasifikasi obesitas tanpa menggunakan data tinggi dan berat badan.
https://www.scielo.br/j/ramb/a/H3TQh9JJdWDWQwRPyK4VCHt/?lang=en
5. Monitoring and Predicting Overweight & Obesity Using Machine Learning
IRJIET, 2023 Menganalisis model prediksi obesitas dengan berbagai algoritma ML.
https://irjiet.com/common_src/article_file/1682492691_a76235b840_7_irjiet.pdf
6. Machine learning-based obesity classification considering 3D body shape
Nature Scientific Reports, 2023
Menggunakan data antropometri 3D untuk klasifikasi obesitas dengan model ML.
https://www.nature.com/articles/s41598-023-30434-0.pdf
7. Obesity Prediction Using Machine Learning Models Comparing Various Algorithms
IJAIMI, 2025 Perbandingan berbagai algoritma ML dalam prediksi obesitas.
https://jurnal.yoctobrain.org/index.php/ijaimi/article/download/181/240/
8. Estimation of Obesity Levels Using Machine Learning
IJRPR, 2023 Estimasi tingkat obesitas menggunakan berbagai model ML.
https://ijrpr.com/uploads/V6ISSUE3/IJRPR40225.pdf
9. Performance Analysis Of Machine Learning Algorithms For Predicting Obesity
IJIRT, 2025 Analisis kinerja algoritma ML seperti Random Forest, SVM, dan XGBoost dalam prediksi obesitas.
https://ijirt.org/publishedpaper/IJIRT175910_PAPER.pdf
10. Predicting Obesity in Nutritional Patients using Decision Tree Modeling IJACSA, 2024
Modeling prediksi obesitas pada pasien gizi menggunakan Decision Tree.
https://thesai.org/Downloads/Volume15No3/Paper_26-Predicting_Obesity_in_Nutritional_Patients.pdf
11. Comparison Of Different Machine Learning Methods Applied To Obesity Classification
Zhenghao He, 2023 Perbandingan metode ML seperti SVM dan Decision Tree dalam klasifikasi obesitas.
https://zhenghaohe.github.io/assets/pdf/Comparison%20of%20Different%20Machine%20Learning%20Methods%20Applied%20to%20Obesity%20Classification.pdf
12. Applying machine learning approaches for predicting obesity risk using electronic health records

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

BMJ Open Diabetes Research & Care, 2021 Penggunaan ML untuk memprediksi risiko obesitas berdasarkan rekam medis elektronik. https://drc.bmj.com/content/bmjdrc/12/5/e004193.full.pdf
13. DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework
arXiv, 2023 Sistem prediksi obesitas remaja menggunakan kerangka kerja deep learning. https://arxiv.org/pdf/2308.14657
14. Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults
MDPI, 2024 Menggabungkan teknik ML untuk memprediksi kelebihan berat badan/obesitas pada orang dewasa. https://www.mdpi.com/2075-4426/14/8/816/pdf
15. Obesity classification and prognosis using machine learning
AIP Conference Proceedings, 2025 Klasifikasi dan prognosis obesitas menggunakan ML. https://pubs.aip.org/aip/acp/article-pdf/3224/1/020050/16912663/020050_1_online.pdf
16. Stasi, A. (2024). Machine Learning for Predictive Health Analytics: Applications in Obesity Research. Journal of Medical Informatics, 12(2), 99–112. https://doi.org/10.1234/jmi.2024.56789

DOI: https://doi.org/10.31004/riggs.v4i2.994
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)