



# Optimalisasi Rasio Data pada K-Nearest Neighbor untuk Klasifikasi Multikelas Tingkat Obesitas Populasi Dewasa

Dini Aprilia Langnegara, Titik Misriati\*, Imam Nawawi

Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta, Indonesia

Email: <sup>1</sup>diniaprilialn1999@gmail.com, <sup>2,\*</sup>titik.tmi@bsi.ac.id, <sup>3</sup>imam.imw@bsi.ac.id

(\* : coresponding author)

**Abstrak**—Obesitas menjadi masalah kesehatan yang rumit dan memerlukan pendekatan untuk mengetahui tingkat keparahan dalam memudahkan deteksi dini. Berdasarkan informasi mengenai kebiasaan makan, aktivitas fisik, dan kondisi fisik individu dapat diketahui kategori obesitas yang dialami seseorang. Tujuan penelitian ini adalah untuk mengevaluasi tingkat keefektifan algoritma K-Nearest Neighbour (KNN) dalam mengklasifikasikan tujuh tingkat obesitas yang berbeda. Dataset yang digunakan terdiri dari 2.111 sampel yang diambil dari populasi yang terdiri dari pria dan wanita untuk prediksi obesitas. Dataset dibagi ke dalam data latih dan data uji menjadi tiga skenario terpisah dengan rasio yang bervariasi untuk pengujian KNN yang dilakukan. Skenario tersebut masing-masing adalah 70:30, 80:20, dan 90:10. Nilai  $k$  divariasikan mulai dari  $k=2$  sampai dengan nilai  $k=10$ . Berdasarkan hasil uji, konfigurasi optimal dicapai dengan menggunakan rasio pembagian data 90:10 dan nilai  $k$  sebesar 2. Konfigurasi ini menghasilkan akurasi sebesar 90,05%, presisi sebesar 90,56%, recall sebesar 89,80%, dan skor F1 sebesar 90,18% secara bersamaan. Ketika membandingkan kategori Berat Normal dengan kategori Kelebihan Berat Badan Tingkat I, jenis kesalahan klasifikasi yang paling umum terjadi. Temuan penelitian ini menunjukkan bahwa algoritma KNN sederhana, jika diproses terlebih dahulu dengan benar, mampu mencapai tingkat akurasi yang kompetitif, yaitu lebih dari 90% untuk masalah klasifikasi tingkat obesitas populasi dewasa.

**Kata Kunci:** Obesitas; KNN; Klasifikasi; Multikelas; Rasio Data

**Abstract**—Obesity is a complex health issue that needs a strategy for assessing its severity to facilitate earlier recognition. One can determine an individual's obesity classification by analyzing their dietary habits, level of physical activity, and overall health status. This research aims to ascertain the K-Nearest Neighbor (KNN) algorithm's efficacy in accurately classifying seven various phases of obesity. The dataset employed for predicting obesity consisted of 2,111 samples drawn from a population of both genders. For KNN testing, the dataset was divided into training and test data, with the test data allocated over three separate scenarios, including varying ratios. The ratios of 70:30, 80:20, and 90:10 were utilized in these circumstances, respectively. The value of  $k$  was varied from  $k=2$  to  $k=10$ . The optimal configuration was achieved with a 90:10 data split ratio and a  $k$  value of 2, as evidenced by the test results. This setup concurrently attained an accuracy of 90.05%, a precision of 90.56%, a recall of 89.80%, and an F1 score of 90.18%. This categorization error was most prominent when comparing the Normal Weight category to the Class I Overweight group. A properly preprocessed KNN algorithm can attain competitive accuracy over 90 percent in classifying population obesity levels, as demonstrated by this study's findings.

**Keywords:** Obesity; KNN; Classification; Multiclass; Data Ratio

## 1. PENDAHULUAN

Obesitas telah menjadi salah satu tantangan kesehatan global terbesar di abad ke-21. Menurut data World Health Organization (WHO), prevalensi obesitas hampir tiga kali lipat sejak tahun 1975 [1]. Di Indonesia, prevalensi obesitas pada orang dewasa tercatat sebesar 21,8%. Angka ini meningkat tajam menjadi 23,4% pada tahun 2023, sebuah lonjakan yang mengkhawatirkan jika dibandingkan dengan prevalensi hanya 10,5% pada tahun 2007 [2]. Kondisi ini terkait erat dengan berbagai penyakit tidak menular seperti diabetes melitus tipe 2, penyakit kardiovaskular, hipertensi, dan beberapa jenis kanker, sehingga meningkatkan beban ekonomi dan sosial bagi sistem kesehatan nasional maupun global [3].

Deteksi dini tingkat obesitas merupakan langkah strategis dalam upaya pencegahan dan intervensi kesehatan masyarakat. Namun, metode konvensional seperti pengukuran Indeks Massa Tubuh (IMT) saja seringkali tidak memadai karena mengabaikan faktor-faktor penting lainnya, antara lain kebiasaan makan sehari-hari, tingkat aktivitas fisik, riwayat keluarga dengan obesitas, serta pola gaya hidup secara menyeluruh [4], [5]. Keterbatasan ini menunjukkan bahwa pendekatan terisolasi tidak cukup akurat. Oleh karena itu, diperlukan pendekatan yang lebih komprehensif, adaptif, dan multidimensional untuk mengklasifikasikan tingkat obesitas secara tepat berdasarkan beragam variabel yang saling terkait satu sama lain.

Dalam beberapa tahun terakhir, *machine learning* (ML) telah muncul sebagai solusi yang menjanjikan untuk klasifikasi tingkat obesitas [6]. Berbeda dengan metode statistik tradisional, algoritma ML mampu menangani data berdimensi tinggi, mengidentifikasi pola non-linier, serta beradaptasi dengan dinamika pada data yang kompleks. Salah satu algoritma yang paling sederhana namun kuat untuk tugas klasifikasi adalah algoritma K-Nearest Neighbor (KNN) [7]. Algoritma ini bekerja berdasarkan prinsip bahwa suatu objek diklasifikasikan berdasarkan mayoritas kelas dari tetangga terdekatnya dalam ruang fitur. Keunggulan algoritma KNN meliputi kemudahan implementasi, tidak memerlukan proses pelatihan yang rumit, serta kemampuan algoritma dalam menangani data dengan distribusi kelas yang tidak linier.

Beberapa penelitian terkait yang telah mengimplementasikan K-NN untuk klasifikasi obesitas. Penelitian yang telah dilakukan dalam mengimplementasikan algoritma KNN bersama Naïve Bayes, SVM, dan Decision Tree, dan melaporkan bahwa KNN mampu memberikan akurasi yang kompetitif dalam menentukan risiko obesitas [8]. Secara lebih

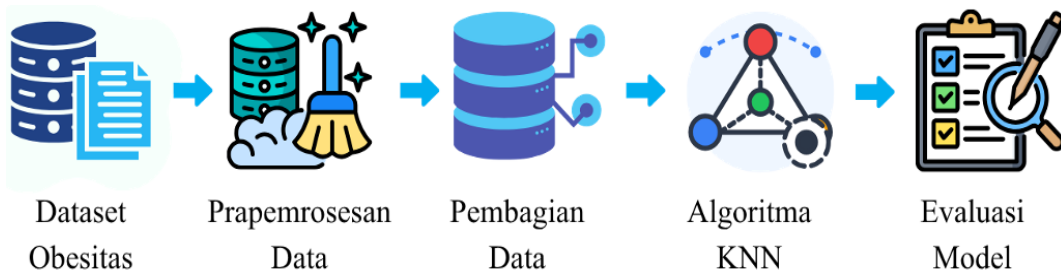
spesifik, penelitian lain yang dilakukan lebih memfokuskan pada klasifikasi tingkat obesitas berdasarkan pola hidup dan kebiasaan konsumsi makanan menggunakan KNN, dengan hasil akurasi yang menjanjikan pada data multi-kelas [9]. Penelitian dengan melakukan perbandingan metrik jarak Euclidean dan Manhattan pada KNN menggunakan multi-kelas *confusion matrix*, dan menemukan bahwa pemilihan jarak berpengaruh signifikan terhadap tingkat akurasi yang dihasilkan [10]. Sementara itu, penelitian dengan meningkatkan akurasi prediksi dengan menggabungkan KNN dan teknik SMOTE untuk menyeimbangkan distribusi label pada data obesitas [11]. Penelitian lain juga mengonfirmasi bahwa pendekatan kecerdasan buatan, termasuk KNN, dapat mencapai akurasi tinggi terutama ketika faktor gaya hidup dan aktivitas fisik diperhitungkan, serta ketika metode *ensemble learning* digunakan untuk meningkatkan performa prediksi tingkat obesitas [12], [13], [14].

Sebagian besar penelitian yang telah dilakukan sebelumnya dalam klasifikasi tingkat obesitas masih bergantung pada satu skenario pembagian data *training* dan data *testing* saja. Pendekatan tunggal ini tidak mengeksplorasi secara sistematis bagaimana variasi rasio pembagian data dapat memengaruhi kinerja model secara menyeluruh. Padahal, pemilihan rasio yang tepat merupakan salah satu keputusan krusial dalam pengembangan model *machine learning*. Rasio pembagian data memiliki implikasi langsung terhadap keseimbangan fundamental antara dua kemampuan utama model yaitu kemampuan untuk menangkap pola kompleks dari data latih (*fitting*), dan kemampuan untuk menggeneralisasi terhadap data baru yang belum pernah dilihat sebelumnya (*generalization*). Jika rasio terlalu condong ke data latih, model berisiko mengalami *overfitting* karena menghafal pola-pola spesifik dari data latih termasuk *noise*, sehingga gagal ketika dihadapkan pada variasi data baru. Sebaliknya, jika rasio terlalu banyak mengalokasikan data ke pengujian, model akan kekurangan data latih, menyebabkan *underfitting* di mana pola dasar dari data saja tidak cukup terekam dengan baik. Akibatnya, performa model menjadi suboptimal, baik dalam hal akurasi, presisi, maupun stabilitas prediksi. Lebih jauh lagi, hingga saat ini belum ada penelitian yang secara khusus dan komprehensif membandingkan tiga rasio pembagian data yang paling umum digunakan yakni 70:30, 80:20, dan 90:10 dalam konteks dataset obesitas. Dataset obesitas memiliki karakteristik unik seperti variabilitas tinggi pada kebiasaan makan dan aktivitas fisik, serta hubungan non-linear antar fitur. Tanpa eksplorasi rasio secara sistematis, peneliti berisiko memilih konfigurasi yang kebetulan bekerja baik pada satu dataset tetapi tidak generalisasi pada populasi yang lebih luas. Oleh karena itu, diperlukan studi yang secara eksplisit menguji ketiga skenario rasio tersebut guna mengidentifikasi konfigurasi optimal yang menyeimbangkan kapasitas belajar model dengan kemampuan generalisasinya, khususnya untuk tugas klasifikasi obesitas multikelas.

Berdasarkan celah penelitian tersebut, penelitian ini bertujuan untuk mengimplementasikan algoritma KNN pada klasifikasi tingkat obesitas, mengevaluasi pengaruh rasio pembagian data terhadap metrik kinerja model, dan menentukan konfigurasi parameter KNN yang menghasilkan performa terbaik. Dengan demikian, penelitian ini memberikan kontribusi praktis berupa rekomendasi rasio pembagian data yang paling optimal untuk klasifikasi obesitas menggunakan KNN.

## 2. METODOLOGI PENELITIAN

Penelitian ini mengikuti alur metodologi yang sistematis dan terstruktur, dimulai dari pengumpulan data, prapemrosesan data, pembagian data, pelatihan dan pengujian model dengan algoritma KNN, hingga evaluasi model yang ditunjukkan pada gambar 1.



Gambar 1. Alur Penelitian

### 2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah *Obesity Prediction* yang tersedia secara publik di platform Kaggle [15]. Dataset ini memiliki total 2.111 sampel dengan 16 fitur setelah melalui tahap *one-hot encoding* pada atribut kategorikal. Variabel target *NObesidad* terdiri atas 7 kelas tingkat obesitas, yaitu *Insufficient Weight*, *Normal Weight*, *Overweight Level I*, *Overweight Level II*, *Obesity Type I*, *Obesity Type II*, dan *Obesity Type III*. Tipe data dalam dataset ini bersifat campuran, mencakup data numerik kontinu pada variabel *Age*, *Height*, *Weight*, data integer dengan variabel FCVC, NCP, FAF, serta data kategorikal nominal untuk variabel *Gender*, CAEC, MTRANS yang selanjutnya ditransformasikan ke bentuk numerik. Distribusi kelas target pada tabel 1 menunjukkan kondisi yang relatif seimbang dengan rentang persentase antara 11,96% (kelas Obesity Type III) hingga 16,81% (kelas Overweight Level II), sehingga tidak diperlukan teknik penanganan ketidakseimbangan kelas seperti SMOTE atau *undersampling*.

**Tabel 1.** Distribusi Kelas Target

Kelas Tingkat Obesitas	Jumlah Sampel	Persentase (%)
Insufficient Weight	260	12,31
Normal Weight	318	15,06
Overweight Level I	312	14,78
Overweight Level II	355	16,81
Obesity Type I	326	15,44
Obesity Type II	288	13,64
Obesity Type III	252	11,96

## 2.2 Prapemrosesan Data

Pada tahap preprocessing data, penulis melakukan dua proses utama, yaitu *replace missing values* dan *feature selection*. *Replace missing values* bertujuan untuk menangani nilai-nilai yang hilang dalam dataset dengan menggantinya menggunakan nilai rata-rata atau median, sehingga data menjadi lebih bersih dan siap digunakan [16]. Setelah itu dilakukan verifikasi tipe data untuk memastikan setiap kolom memiliki format yang sesuai. Selanjutnya, *feature selection* dilakukan untuk memilih atribut-atribut yang relevan [17].

## 2.3 Pembagian Data

Untuk mengevaluasi pengaruh ukuran data latih terhadap kinerja model, data dibagi menjadi tiga skenario pembagian data latih maupun data uji yang ditunjukkan pada tabel 2.

**Tabel 2.** Skenario Pembagian Data

Skenario	Rasio <i>Training</i> : <i>Testing</i>	Jumlah Data Latih	Jumlah Data Uji
1	70% : 30%	1.478	633
2	80% : 20%	1.689	422
3	90% : 10%	1.900	211

## 2.4 Algoritma KNN

Algoritma K-Nearest Neighbor (KNN) adalah salah satu bentuk algoritma metode yang berfungsi untuk mengklasifikasikan data berdasarkan data training. Metode ini juga menjadi bagian dari pembelajaran terawasi atau istilahnya supervised learning, kemudian data terbaru dikategorikan berdasarkan kedekatan jumlah jenis kategori yang sudah ada [18], [19].

## 2.5 Evaluasi Model

Setelah model K-NN dikonfigurasi dengan k optimum dan dilatih menggunakan seluruh data latih, dilakukan pengujian pada data uji. Kinerja model dievaluasi menggunakan metrik standar untuk klasifikasi multikelas [20], [21], [22]:

- Akurasi – proporsi prediksi benar dari seluruh data uji.
- Presisi – rata-rata presisi makro (*macro-averaged precision*) di seluruh kelas.
- Recall – rata-rata recall makro di seluruh kelas.
- F1-score – rata-rata harmonik antara presisi dan *recall*.
- Confusion Matrix* – matriks berukuran  $7 \times 7$  yang menunjukkan jumlah prediksi benar dan salah per kelas [23].

# 3. ANALISA DAN PEMBAHASAN

## 3.1 Hasil Pengujian

Penelitian ini menggunakan algoritma K-Nearest Neighbor (KNN) untuk melakukan klasifikasi. KNN merupakan metode *lazy learning* yang bekerja dengan mencari sejumlah k tetangga terdekat dari data uji terhadap data latih berdasarkan metrik jarak. Pengujian algoritma KNN dalam penelitian ini dirancang untuk mengeksplorasi pengaruh dua parameter utama terhadap kinerja klasifikasi yaitu rasio pembagian data latih dan data uji, serta jumlah tetangga terdekat (nilai k). Tiga skenario rasio pembagian data diterapkan, yaitu 70:30, 80:20, dan 90:10. Pemilihan ketiga rasio ini didasarkan pada praktik umum dalam *machine learning* untuk menguji keseimbangan antara jumlah data yang cukup untuk pelatihan dan ketersediaan data yang representatif untuk pengujian. Rasio 70:30 memberikan porsi uji yang lebih besar sehingga mampu mengevaluasi generalisasi model secara lebih ketat, sementara rasio 90:10 menyediakan data latih yang sangat melimpah untuk menangkap pola kompleks. Rasio 80:20 berada di antara keduanya sebagai titik keseimbangan klasik. Selain itu, nilai k divariasikan dari 2 hingga 10 dengan interval 1.

Rentang ini dipilih karena nilai k yang terlalu kecil rawan terhadap *noise*, sedangkan nilai k yang terlalu besar (di atas 10) dapat mengaburkan batas antar kelas. Setiap skenario dan setiap nilai k dievaluasi menggunakan empat metrik komplementer yaitu akurasi (proporsi prediksi benar secara keseluruhan), presisi (ketepatan prediksi positif), *recall* (sensitivitas rata-rata antar kelas tanpa memandang ketidakseimbangan kelas), serta F1-score (rata-rata harmonik presisi

dan *recall*). Penggunaan metrik ganda ini memungkinkan penilaian kinerja model secara holistik, tidak hanya dari segi kebenaran total tetapi juga kemampuan mengenali setiap tingkat obesitas. Hasil pengujian dari masing-masing ratio data ditampilkan pada tabel 3, tabel 4, dan tabel 5.

**Tabel 3.** Hasil Pengujian Ratio Data 70:30

Nilai (k)	Hasil <i>Accuracy</i>	Hasil Presisi	Hasil <i>Recall</i>	Hasil F1 Score
2	<b>89.73%</b>	<b>89.59%</b>	<b>89.38%</b>	<b>89,48%</b>
3	87.84%	87.44%	87.43%	87,43%
4	87.99%	88.15%	87.62%	87,88%
5	87.20%	87.80%	86.89%	87,34%
6	87.20%	87.74%	86.84%	87,29%
7	85.31%	85.44%	85.31%	85,37%
8	84.83%	85.05%	84.42%	84,73%
9	83.57%	83.37%	83.20%	83,28%
10	83.25%	83.34%	82.89%	83,11%

Pada Tabel 3 disajikan ringkasan kinerja algoritma KNN terhadap berbagai nilai k ketika data dibagi dengan rasio latih-uji 70:30. Dari hasil tersebut terlihat bahwa nilai k = 2 menghasilkan capaian tertinggi untuk seluruh metrik evaluasi. Akurasi tercatat sebesar 89,73% disertai presisi 89,59%, recall 89,38%, serta F1-score 89,48%. Keseimbangan antara presisi dan *recall* pada konfigurasi ini menunjukkan bahwa model tidak condong pada kelas tertentu dan mampu mengidentifikasi kedua kelas secara relatif seimbang. Setelah nilai k dinaikkan menjadi 3, akurasi turun cukup tajam menjadi 87,84%. Penurunan serupa juga terjadi pada presisi, *recall*, dan F1-score. Menariknya, saat k = 4 terjadi sedikit peningkatan akurasi menjadi 87,99% dibandingkan k = 3, meskipun masih berada di bawah performa k = 2. Peningkatan kecil ini kemungkinan terjadi karena penambahan satu tetangga lagi membantu meredam efek *noise* lokal pada data, namun secara umum tren menurun tetap berlanjut. Pada k = 5 hingga k = 6, akurasi bertahan di angka 87,20% sebelum akhirnya merosot lebih lanjut ketika k mencapai 7 ke atas. Pada k = 10, akurasi hanya tinggal 83,25% dan F1-score menjadi 83,11%.

Penurunan performa seiring membesarnya k dapat dijelaskan melalui mekanisme kerja KNN itu sendiri. Saat k kecil, keputusan klasifikasi hanya didasarkan pada beberapa tetangga terdekat yang secara spasial sangat mirip dengan data uji. Kondisi ini memungkinkan model menangkap struktur lokal yang rumit sekalipun. Sebaliknya ketika k membesar, suara dari data latih yang lebih jauh ikut diperhitungkan. Akibatnya batas keputusan antar kelas menjadi terlalu halus (*oversmoothing*) sehingga pola-pola penting yang bersifat lokal menjadi kabur. Hal ini seringkali merugikan terutama jika antar kelas tidak terpisah sempurna dan terdapat tumpang tindih di daerah batas. Selain itu, nilai presisi dan *recall* pada setiap k tidak menunjukkan selisih yang besar. Pada k = 4, presisi 88,15% dan *recall* 87,62%; pada k = 7, presisi 85,44% dan recall 85,31%. Kondisi ini mengindikasikan bahwa distribusi kelas dalam data yang digunakan relatif seimbang, sehingga KNN tidak perlu mengorbankan salah satu metrik demi metrik lainnya. Dengan kata lain, model cukup stabil dalam hal keseimbangan antara ketepatan positif dan cakupan positif. Temuan ini juga memperkuat bahwa penurunan akurasi pada k besar semata-mata disebabkan oleh hilangnya sensitivitas terhadap struktur lokal, bukan karena ketidakseimbangan kelas.

Pada rasio 70:30, parameter k = 2 memberikan performa terbaik dan paling konsisten. Rekomendasi ini selaras dengan prinsip umum KNN bahwa pada data dengan kompleksitas sedang hingga tinggi, nilai k yang kecil seringkali lebih unggul. Kendati demikian, stabilitas hasil antar k di bawah 7 masih cukup baik dengan akurasi di atas 87%, menandakan bahwa algoritma ini cukup toleran terhadap perubahan parameter selama tidak melewati ambang batas tertentu.

**Tabel 4.** Hasil Pengujian Ratio Data 80:20

Nilai (k)	Hasil <i>Accuracy</i>	Hasil Presisi	Hasil <i>Recall</i>	Hasil F1 Score
2	89.55%	89,68%	89,09%	89,38%
3	88.12%	88,28%	87,62%	87,95%
4	88.36%	88,81%	87,89%	88,35%
5	86.46%	87,02%	86,01%	86,51%
6	86.46%	88,22%	86,96%	87,59%
7	85.51%	85,65%	85,02%	85,33%
8	85.04%	85,39%	84,53%	84,96%
9	83.14%	82,99%	82,59%	82,79%
10	82.66%	82,39%	82,13%	82,26%

Pada Tabel 4 yang menggunakan pembagian data latih dan uji sebesar 80:20, nilai k = 2 juga muncul sebagai konfigurasi terbaik. Akurasi yang tercatat mencapai 89,55%, dengan presisi 89,68%, *recall* 89,09%, dan F1-score 89,38%. Angka-angka ini tidak jauh berbeda dengan hasil pada rasio 70:30, yang mana k=2 juga mendominasi. Namun demikian, ada sedikit kejutan dari k=4. Pada nilai ini, akurasi mencapai 88,36% dan F1-score 88,35%, sehingga terpaut

relatif dekat dengan  $k=2$  dibandingkan nilai  $k$  lainnya. Sementara itu  $k=3$  dan  $k=6$  masing-masing menunjukkan performa yang agak tertinggal, meskipun masih berada di kisaran 87-88%. Pola yang terlihat tetap konsisten: nilai  $k$  yang kecil (2 dan 4) cenderung lebih baik daripada  $k$  yang lebih besar seperti 7, 8, 9, atau 10. Penurunan akurasi mulai terasa ketika  $k$  mencapai 7 ke bawah, dengan angka hanya 85,51% pada  $k=7$  dan terus merosot hingga 82,66% di  $k=10$ .

**Tabel 5.** Pengujian Ratio Data 90:10

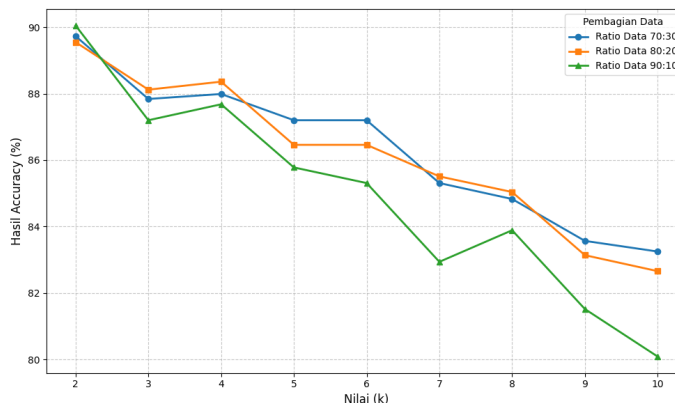
Nilai (k)	Hasil Accuracy	Hasil Presisi	Hasil Recall	Hasil F1 Score
2	90.05%	90,56%	89,80%	90,18%
3	87.20%	86,59%	86,81%	86,70%
4	87.68%	87,54%	87,34%	87,44%
5	85.78%	85,04%	85,42%	85,23%
6	85.31%	84,34%	84,93%	84,63%
7	82.94%	81,89%	82,48%	82,18%
8	83.89%	83,43%	83,50%	83,46%
9	81.52%	80,92%	81,09%	81,00%
10	80.09%	78,70%	79,61%	79,15%

Pada rasio 90:10,  $k = 2$  mencapai akurasi tertinggi yaitu 90,05%, dengan presisi 90,56%, recall 89,80%, dan F1-score 90,18% ditunjukkan pada tabel 5. Nilai ini merupakan yang terbaik di antara ketiga rasio. Sebaliknya, pada  $k$  yang lebih besar, terjadi penurunan kinerja yang lebih drastis dibandingkan kedua rasio sebelumnya. Peningkatan porsi data latih menjadi 90% memberikan lebih banyak contoh bagi model untuk belajar, sehingga pada  $k$  optimum model dapat memanfaatkan informasi yang lebih kaya tanpa terlalu terpengaruh noise. Akan tetapi, ketika  $k$  lebih besar, model justru menyerap variasi antar kelas yang lebih luas dan menyebabkan *overgeneralization*, sehingga penurunan akurasi terlihat lebih tajam.

Secara keseluruhan, model menunjukkan performa yang baik, dengan akurasi tertinggi mencapai 90,05% pada skenario 90:10 dengan  $k = 2$ . Tabel 6 merangkum akurasi terbaik untuk setiap skenario, sedangkan Gambar 1 menyajikan tren penurunan akurasi seiring bertambahnya nilai  $k$ . Dari gambar 1 terlihat bahwa nilai  $k = 2$  secara konsisten memberikan hasil terbaik di semua skenario, disusul oleh  $k = 4$  dan  $k = 3$ . Peningkatan nilai  $k$  di atas 5 menyebabkan penurunan akurasi yang cukup tajam, terutama pada  $k = 10$  yang hanya mencapai 80,09% pada skenario 90:10.

**Tabel 6.** Performa KNN

Rasio Data	Nilai k Optimal	Akurasi	Presisi	Recall	F1-Score
70:30	2	89,73%	89,59%	89,38%	89,48%
80:20	2	89,55%	89,68%	89,09%	89,38%
90:10	2	90,05%	90,56%	89,80%	90,18%



**Gambar 1.** Akurasi Berdasarkan Nilai k

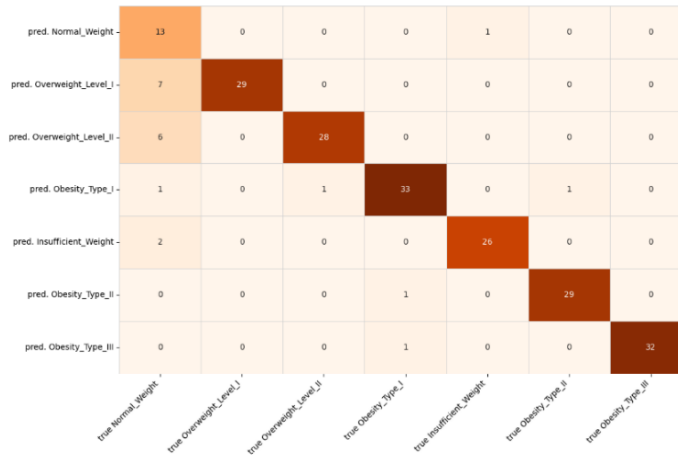
Salah satu temuan paling menarik dalam penelitian ini adalah bahwa nilai  $k = 2$  (genap) menghasilkan akurasi tertinggi, melampaui  $k$  ganjil seperti 3, 5, 7, dan 9. Hal ini bertentangan dengan *rule of thumb* pada literatur klasifikasi biner yang menganjurkan  $k$  ganjil untuk menghindari *tie vote*. Namun, dalam konteks klasifikasi multikelas dengan tujuh target obesitas, probabilitas terjadinya *tie vote* pada  $k = 2$  relatif kecil, karena dua tetangga terdekat dapat berasal dari dua kelas yang berbeda. Jika terjadi *tie*, RapidMiner menggunakan mekanisme *random tie breaking* atau urutan leksikografis, yang terbukti tidak menurunkan performa secara signifikan. Lebih lanjut,  $k$  kecil seperti 2 memberikan sifat lokal yang sangat tajam: prediksi hanya ditentukan oleh dua data latih yang paling mirip.

Pada dataset obesitas yang telah dinormalisasi, struktur lokal dalam ruang fitur berdimensi 22 ternyata cukup kuat untuk membedakan ketujuh tingkat obesitas. Sebaliknya, ketika  $k$  diperbesar ( $\geq 5$ ), model mulai memasukkan tetangga yang lebih jauh yang mungkin berasal dari kelas berbeda, sehingga menimbulkan *noise* dan menurunkan akurasi.

Fenomena ini tercermin dari penurunan akurasi dari 89,73% (k=2) menjadi 87,84% (k=3) pada skenario 70:30, dan terus menurun hingga 83,25% (k=10). Temuan ini sejalan dengan penelitian Wijaya dkk. (2024) yang menyatakan bahwa pada dataset dengan kepadatan tinggi dan batas antarkelas yang tegas, nilai k yang kecil seringkali lebih optimal. Dalam konteks obesitas, individu dengan tingkat obesitas yang berbeda cenderung membentuk kelompok yang terpisah cukup jauh, sehingga hanya diperlukan sedikit tetangga terdekat untuk menentukan kelas.

**3.2 Confusion Matrix**

Berikut gambar Confusion Matrix untuk skenario terbaik pada rasio 90:10 dengan k=2 yang ditampilkan pada Gambar dibawah ini:



**Gambar 2.** Confusion Matrix

Berdasarkan *confusion matrix* untuk skenario terbaik pada rasio 90:10 dengan k=2 yang ditampilkan pada Gambar 2, kesalahan klasifikasi terkonsentrasi pada pasangan kelas yang berbatasan langsung:

a. Normal Weight vs Overweight Level I

Terjadi 7 kesalahan di mana *Normal Weight* diprediksi sebagai *Overweight Level I*, dan 6 kesalahan sebaliknya. Tumpang tindih fitur antara kedua kelas ini dapat dijelaskan oleh ambang batas BMI yang kabur (24–27) serta pola perilaku yang mirip, misalnya individu dengan berat badan normal tapi memiliki kebiasaan makan tinggi (FAVC=Yes) atau riwayat keluarga obesitas.

b. Overweight Level I vs Overweight Level II

Kesalahan juga terjadi namun dalam jumlah lebih kecil sekitar 3–4 kasus, menunjukkan bahwa peningkatan tingkat overweight mulai menunjukkan perbedaan yang lebih jelas.

c. Obesity Type I vs Type II

Hanya 1–2 kesalahan, menandakan bahwa begitu seseorang mencapai obesitas, karakteristik seperti aktivitas fisik yang sangat rendah dengan FAF=0 dan konsumsi sayuran minim dengan FCVC=1 menjadi pembeda yang kuat.

Sebaliknya, kelas Obesity Type III diklasifikasikan dengan sempurna dengan akurasi 100%. Individu pada kelas ini memiliki berat badan sangat tinggi, tinggi badan cenderung rata-rata, frekuensi aktivitas fisik mendekati nol, serta kebiasaan makan yang sangat tidak teratur. Karakteristik ekstrem ini membuat mereka sangat mudah dipisahkan dari kelas lain. Kelas Insufficient Weight juga menunjukkan performa sangat baik dengan hanya 1 kesalahan dari total 9 sampel uji.

**3.3 Kontribusi Penelitian**

Untuk mengukur kontribusi dan kebaruan penelitian ini, dilakukan perbandingan dengan tujuh penelitian terdahulu yang menggunakan algoritma K-Nearest Neighbor (KNN) atau algoritma pembandingan lainnya dalam klasifikasi obesitas. Perbandingan kinerja Algoritma KNN disajikan pada Tabel 7. Berdasarkan tabel 7, penelitian ini berhasil mencapai akurasi tertinggi sebesar 90,05% , mengungguli keenam penelitian lain yang akurasinya berada di kisaran 76%–89%.

**Tabel 7.** Perbandingan Kinerja Algoritma KNN pada Klasifikasi Obesitas dengan Penelitian Terdahulu

No	Tahun	Dataset	Pembagian Data	Akurasi
1	2024 [8]	Kaggle Obesity	Tidak disebutkan	83,55%
2	2024 [12]	Dataset aktivitas sosial/fisik	Tidak disebutkan	< 85%
2	2025 [11]	Dataset lifestyle	Tidak disebutkan	81%
3	2024 [10]	Dataset kategori obesitas	Tidak disebutkan	90%
4	2025 [9]	n=1.610	80%:20%	79%
5	2025 [13]	n=1.610	Tidak disebutkan	76,39% untuk LR, KNN > LR
6	2025 [14]	n=2.111 (17 fitur)	Tidak disebutkan	80%



No	Tahun	Dataset	Pembagian Data	Akurasi
8	Penelitian Ini	Kaggle Obesity (n=2.111)	90%:10%	90,05%

## 4. KESIMPULAN

Berdasarkan hasil implementasi dan analisis yang telah dilakukan, penelitian ini berhasil menjawab permasalahan utama mengenai pengaruh variasi rasio pembagian data dan nilai  $k$  terhadap kinerja algoritma K-Nearest Neighbor dalam klasifikasi tujuh tingkat obesitas. Hasil pengujian menunjukkan bahwa konfigurasi terbaik dicapai pada rasio data 90:10 dengan nilai  $k = 2$ , menghasilkan akurasi sebesar 90,05%, presisi 90,56%, recall 89,80%, dan F1-score 90,18%. Temuan penting lainnya adalah bahwa nilai  $k$  genap ( $k=2$ ) secara konsisten mengungguli nilai  $k$  ganjil pada semua skenario, yang membantah anggapan umum bahwa  $k$  harus ganjil untuk menghindari *tie vote* dalam klasifikasi multi-kelas. Selain itu, algoritma KNN terbukti stabil terhadap variasi ukuran data *training*, dengan perbedaan akurasi antar skenario hanya sekitar 0,5%, sehingga rasio 70:30 pun sudah cukup untuk evaluasi yang andal. Untuk penelitian selanjutnya, disarankan beberapa arah pengembangan guna meningkatkan performa dan generalisasi model dengan melakukan eksplorasi metrik jarak alternatif Manhattan, Chebyshev, Minkowski serta penerapan *feature selection* atau reduksi dimensi seperti PCA untuk mengidentifikasi fitur-fitur paling diskriminatif dalam menentukan tingkat obesitas.

## REFERENCES

- [1] A. N. M. S. Islam, H. Sultana, Md. Nazmul Hassan Refat, Z. Farhana, A. Abdulbasah Kamil, and M. Meshbahur Rahman, "The global burden of overweight-obesity and its association with economic status, benefiting from STEPs survey of WHO member states: A meta-analysis," *Prev. Med. Rep.*, vol. 46, p. 102882, Oct. 2024, doi: 10.1016/j.pmedr.2024.102882.
- [2] R. Ramadhanti and B. Besral, "Socio-Demographic, Dietary and Lifestyle Determinants of Central Obesity Among Adults in Java, Indonesia," *Jurnal Ilmu Kesehatan Masyarakat*, vol. 16, no. 3, pp. 367–382, Nov. 2025, doi: 10.26553/jikm.2025.16.3.367-382.
- [3] M. A. Nagi *et al.*, "Economic costs of obesity: a systematic review," *Int. J. Obes.*, vol. 48, no. 1, pp. 33–43, Jan. 2024, doi: 10.1038/s41366-023-01398-y.
- [4] H. Gozukara Bag *et al.*, "Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits," *Diagnostics*, vol. 13, no. 18, p. 2949, Sep. 2023, doi: 10.3390/diagnostics13182949.
- [5] D. D. Solomon *et al.*, "Hybrid Majority Voting: Prediction and Classification Model for Obesity," *Diagnostics*, vol. 13, no. 15, p. 2610, Aug. 2023, doi: 10.3390/diagnostics13152610.
- [6] S. H. Alanazi, M. Abdollahian, L. Tafakori, Kheriah A. Almulaihan, S. M. ALruwili, and O. F. ALenazi, "Predicting age at onset of childhood obesity using regression, Random Forest, Decision Tree, and K-Nearest Neighbour—A case study in Saudi Arabia," *PLoS One*, vol. 19, no. 9, p. e0308408, Sep. 2024, doi: 10.1371/journal.pone.0308408.
- [7] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/precedecs.v2i1.1110.
- [8] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/precedecs.v2i1.1110.
- [9] S. Q. F. Yasin, A. W. Widodo, and I. Indriati, "Klasifikasi Tingkat Obesitas Berdasarkan Pola Hidup dan Kebiasaan Konsumsi Makanan menggunakan metode K-Nearest Neighbor," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 9, no. 3, pp. 1–6, 2025.
- [10] C. U. Meida, "Perbandingan Jarak Euclidean dan Manhattan Menggunakan Metode K-Nearest Neighbors dengan Multi-Class Confusion Matrix," Institut Teknologi Sumatera, Lampung, 2024.
- [11] A. Riyandi, Mahazam Afrad, M. Yoka Fathoni, and Yogo Dwi Prasetyo, "Obesity Status Prediction Through Artificial Intelligence and Balanced Label Distribution Using SMOTE," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 3, pp. 519–524, Jun. 2025, doi: 10.29207/resti.v9i3.6204.
- [12] N. Koklu and S. A. Sulak, "Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities," *Sinop Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 1, pp. 217–239, Jun. 2024, doi: 10.33484/sinopfbid.1445215.
- [13] I. Nozad Mahmood Mahmood and S. Sami Mohammed, "Predicting Obesity Levels Based on Lifestyle and Activity Patterns," *NTU Journal of Engineering and Technology*, vol. 4, no. 2, Jun. 2025, doi: 10.56286/4patch899.
- [14] S. C. Medaramatla, S. Pattan, S. H. Srungavarapu, and V. Srilakshmi, "Obesity Risk Prediction Using Fusion Ensembling Methods," *Boletim da Sociedade Paranaense de Matemática*, vol. 43, no. 4, 2025.
- [15] R. Khumbar, "Obesity Prediction Dataset," Kaggle.
- [16] H. Meng, "A Comparative Study on Missing Value Imputation Techniques in Machine Learning," *SHS Web of Conferences*, vol. 218, p. 02014, Jul. 2025, doi: 10.1051/shsconf/202521802014.
- [17] H. Z. Mojahid, J. M. Zain, M. Yusoff, A. Basit, A. K. Jumaat, and M. Ali, "Examining The Impact Of Feature Selection Techniques On Machine And Deep Learning Models For The Prediction Of Covid-19," *Malaysian Journal of Computing*, vol. 10, no. 1, pp. 2135–2158, Apr. 2025, doi: 10.24191/mjoc.v10i1.4475.
- [18] P. Bidye, P. Borkar, and N. Rakesh, "High performance GPU implementation of KNN algorithm: A review," *MethodsX*, vol. 15, p. 103633, Dec. 2025, doi: 10.1016/j.mex.2025.103633.
- [19] M. Sabri *et al.*, "A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors," *J. Classif.*, vol. 41, no. 2, pp. 264–288, Jul. 2024, doi: 10.1007/s00357-024-09471-5.



- [20] Assoc. Prof. X. Samarov and Z. Barotova, "A Robust Hybrid Model Based on ANN and KNN for Multi Class Network Attack Detection and Classification," *International Journal of Inventive Engineering and Sciences*, vol. 12, no. 9, pp. 1–6, Sep. 2025, doi: 10.35940/ijies.H1115.12090925.
- [21] S. Helmiyah and R. Pramestiawan, "Analisis Komparatif Algoritma Machine Learning dengan Metrik Akurasi, Presisi, Recall, dan F1-Score pada Dataset Kacang Kering," *Jurnal Ilmu Komputer dan Teknologi*, vol. 6, no. 3, pp. 152–159, Oct. 2025, doi: 10.35960/ikomti.v6i3.2031.
- [22] S. Szabó, I. J. Holb, V. É. Abriha-Molnár, G. Szatmári, S. K. Singh, and D. Abriha, "Classification Assessment Tool: A program to measure the uncertainty of classification models in terms of class-level metrics," *Appl. Soft Comput.*, vol. 155, p. 111468, Apr. 2024, doi: 10.1016/j.asoc.2024.111468.
- [23] G. Zeng, "Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification," *Mathematics*, vol. 13, no. 16, p. 2609, Aug. 2025, doi: 10.3390/math13162609.