# Analisis Perbandingan Model Decision Tree dan K-Nearest Neighbors (KNN) untuk Estimasi Body Fat Percentage

**Sinta Khoirinnisa*[1], Siti Ulya Ainur Rohmah[2], Mohammad Ali Fikri[3], Wahyu Prastyo Hariyadi[4], and Mohammad Nur Fawaiq[5]**

**1-5 Universitas YPPI Rembang**
**Jl. Raya Rembang – Pamotan KM 4 Rembang, Jawa Tengah**
**sintakhoirinnisa@gmail.com; ulyaainur26@gmail.com; wibugamer30@gmail.com; prasetyowahyu@gmail.com; mohnurfawaiq@gmail.com**

## Abstrak

Persentase lemak tubuh atau Body Fat Percentage (BFP) merupakan parameter penting untuk mengevaluasi kondisi kesehatan dan tingkat obesitas. Namun, pengukuran BFP secara langsung membutuhkan perangkat khusus seperti BodPod atau DEXA yang tidak selalu tersedia dan memerlukan biaya tinggi. Oleh karena itu, prediksi BFP berbasis machine learning menjadi alternatif yang lebih efisien. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma regresi, yaitu Decision Tree Regression dan K-Nearest Neighbors (KNN) Regression, dalam memprediksi BFP berdasarkan variabel antropometri. Dataset yang digunakan terdiri dari 252 sampel dengan 14 fitur seperti density, umur, berat badan, tinggi badan, lingkar perut, pinggul, dan lainnya. Evaluasi model menggunakan metrik MAE, MSE, RMSE, dan $R^2$. Hasil penelitian menunjukkan bahwa Decision Tree memiliki performa lebih baik dengan MAE 0,5627, RMSE 1,4049, dan $R^2$ sebesar 0,9575, sedangkan **KNN** menghasilkan MAE 2,4937, RMSE 3,0547, dan $R^2$ sebesar 0,7994. Dengan demikian, Decision Tree lebih direkomendasikan untuk estimasi BFP menggunakan data antropometri.

**Kata Kunci** Body Fat, Decision Tree, KNN, Machine Learning, Antropometri

**Digital Object Identifier** 10.36802/jnanaloka.2026.v7-no1-1-9

## 1 Pendahuluan

Persentase lemak tubuh atau *Body Fat Percentage* (BFP) merupakan indikator penting yang digunakan untuk menggambarkan proporsi jaringan lemak dalam tubuh seseorang [1]. Parameter ini memiliki peranan besar dalam berbagai bidang, mulai dari kesehatan, kebugaran, hingga penelitian klinis. Selain menjadi penanda tingkat obesitas, BFP juga berhubungan dengan potensi munculnya penyakit kronis seperti diabetes, hipertensi, dislipidemia, dan penyakit jantung koroner. Karena itu, pengukuran BFP secara akurat menjadi kebutuhan penting dalam penilaian kondisi kesehatan seseorang [2].

Berbagai metode telah dikembangkan untuk mengukur tingkat lemak tubuh secara langsung, seperti *Dual-Energy X-ray Absorptiometry* (DEXA), densitometri, dan *hydrostatic weighing* yang dinilai memiliki tingkat akurasi tinggi [3]. Namun, penggunaan metode tersebut tidak selalu mudah diaplikasikan [4]. Selain memerlukan biaya pemeriksaan yang cukup tinggi, alat ukur ini hanya tersedia di fasilitas tertentu dan harus dioperasikan oleh tenaga

* Corresponding author.

profesional. Kondisi ini membuat metode kurang efisien untuk digunakan pada populasi besar atau dalam pengukuran rutin.

Sebagai alternatif, data antropometri mulai banyak dimanfaatkan karena lebih mudah diperoleh dan tidak membutuhkan peralatan mahal. Data antropometri seperti lingkar perut, berat badan, tinggi badan, lingkar pinggul, dan ukuran tubuh lainnya dapat memberikan gambaran awal terhadap komposisi tubuh [3]. Namun, untuk menghasilkan estimasi BFP yang akurat dari data antropometri, diperlukan teknik analisis yang mampu menangani hubungan variabel yang kompleks. Oleh sebab itu, pendekatan *machine learning* menjadi relevan, karena mampu mempelajari pola data secara lebih mendalam dan adaptif [5].

Perkembangan teknologi *machine learning* memungkinkan berbagai model digunakan untuk memprediksi parameter kesehatan, termasuk BFP [6]. Algoritma regresi dapat menangkap hubungan linear maupun nonlinear sehingga dapat memberikan estimasi yang lebih mendekati pengukuran langsung. Walaupun demikian, performa setiap algoritma sangat dipengaruhi oleh struktur data, jumlah sampel, serta karakteristik variabel yang digunakan [7]. Oleh karena itu, pemilihan model yang tepat memerlukan proses evaluasi dan perbandingan.

Dalam penelitian ini, dua algoritma *machine learning* dipilih untuk dibandingkan, yaitu Decision Tree Regression dan K-Nearest Neighbors Regression (KNN) [8]. Decision Tree merupakan algoritma yang bekerja dengan memecah data menjadi beberapa cabang berdasarkan nilai atribut tertentu sehingga mampu menangkap pola nonlinear kompleks dengan mudah [9]. Algoritma ini juga memiliki kelebihan berupa interpretasi model yang jelas dan tidak memerlukan proses normalisasi data. Sementara itu, KNN memprediksi nilai berdasarkan kedekatan antar data berdasarkan jarak, namun performanya sangat bergantung pada skala serta distribusi sampel [10].

Pemilihan kedua algoritma tersebut didasarkan pada popularitas dan kesederhanaannya dalam pemodelan regresi. Decision Tree dan KNN sering digunakan dalam penelitian yang melibatkan data numerik, terutama dalam konteks prediksi biometrik atau variabel medis [11]. Penelitian [8] membandingkan kedua algoritma untuk klasifikasi tingkat obesitas dan menemukan bahwa Decision Tree unggul dalam akurasi. Penelitian [11] juga menunjukkan bahwa Decision Tree memiliki performa lebih baik daripada KNN untuk prediksi stroke awal. Sementara itu, penelitian [12] menggunakan tiga model termasuk Decision Tree untuk estimasi BFP, namun tidak mengikutsertakan KNN sebagai pembanding. Meskipun kedua algoritma termasuk dalam kategori metode sederhana, karakteristiknya berbeda secara fundamental. Oleh karena itu, membandingkan kedua algoritma ini penting dilakukan untuk mengetahui mana yang lebih efektif dalam memprediksi BFP secara akurat pada dataset antropometri yang tersedia [13].

Secara keseluruhan, tujuan dari penelitian ini adalah menganalisis dan membandingkan kinerja Decision Tree Regression dan KNN Regression dalam memprediksi Body Fat Percentage [12]. Dengan menggunakan dataset antropometri yang terdiri dari 252 sampel dan 14 fitur fisik, penelitian ini menilai performa kedua algoritma menggunakan metrik evaluasi seperti MAE, MSE, RMSE, dan nilai koefisien determinasi $R^2$ [14]. Hasil penelitian ini diharapkan dapat memberikan referensi bagi pengembangan metode prediksi BFP yang lebih praktis dan efisien, terutama pada kondisi di mana alat ukur profesional tidak tersedia [15].

Berdasarkan tinjauan literatur, sejauh pengetahuan penulis, penelitian yang secara khusus membandingkan Decision Tree Regression dan K-Nearest Neighbors Regression untuk estimasi Body Fat Percentage berbasis data antropometri dengan 14 fitur fisik masih sangat terbatas. Penelitian [11] membandingkan kedua algoritma untuk prediksi stroke, namun tidak pada konteks BFP. Penelitian [13] memprediksi obesitas menggunakan beberapa algoritma, tetapi tidak melakukan perbandingan mendalam antara Decision Tree dan KNN

secara spesifik. Sementara itu, penelitian [12] menggunakan tiga model untuk estimasi BFP tetapi tidak mencakup KNN. Oleh karena itu, penelitian ini hadir untuk mengisi celah tersebut dengan memberikan perbandingan kuantitatif yang sistematis menggunakan metrik MAE, MSE, RMSE, dan $R^2$.

# 2 Metodologi

## 2.1 Dataset

Data yang digunakan adalah *Body Fat Prediction Dataset*, diperoleh dari Kaggle (`https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset`). Dataset tersebut berisi 252 data sampel antropometri dari responden laki-laki dewasa dengan 14 variabel antropometri sebagai fitur utama dan satu variabel target berupa BodyFat. Fitur-fitur yang tercatat dalam dataset yaitu *Density, Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist.* Variabel tersebut digunakan sebagai prediktor untuk memperkirakan nilai persentase lemak tubuh. Implementasi dilakukan menggunakan bahasa Python 3.10 dengan *library* scikit-learn 1.2.2, pandas, numpy, dan matplotlib.

## 2.2 Tahapan Penelitian

Langkah-langkah dalam penelitian ini dilakukan secara bertahap sebagai berikut:

1. **Preprocessing data.** Pada tahap awal, data dibersihkan serta dipersiapkan untuk proses pelatihan model. Tahapan ini termasuk:
   a. Memisahkan variabel fitur dan target.
   b. Melakukan normalisasi khusus untuk model KNN.
   c. Membagi dataset menjadi data latih dan data uji.

   Pembagian data dilakukan menggunakan skema *train-test split* dengan perbandingan 80% sebagai data latih dan 20% sebagai data uji.
2. **Training model.** Dua algoritma regresi diterapkan, yaitu:
   a. **Decision Tree Regression.** Hyperparameter Decision Tree yang digunakan antara lain `max_depth=7`, `min_samples_split=5`, dan `criterion='gini'`. Parameter ini diperoleh dari proses *tuning* menggunakan `GridSearchCV` dengan *5-fold cross-validation.*
   b. **K-Nearest Neighbors Regression** ($k = 5$). Nilai $k = 5$ dipilih setelah uji coba dengan rentang $k = 3$ hingga $k = 11$, dimana $k = 5$ memberikan nilai RMSE terendah pada data validasi.

   Masing-masing model dilatih menggunakan data latih yang telah diproses.
3. **Evaluasi model.** Untuk mengukur kinerja kedua algoritma, digunakan beberapa metrik evaluasi yaitu:
   a. *Mean Absolute Error* (MAE)
   b. *Mean Squared Error* (MSE)
   c. *Root Mean Squared Error* (RMSE)
   d. Koefisien Determinasi ($R^2$)
4. **Analisis hasil.** Hasil evaluasi kedua model dianalisis untuk melihat model mana yang memberikan performa paling efektif dalam memprediksi Body Fat Percentage.

# 3 Hasil dan Pembahasan

## 3.1 Tampilan Dataset

Dataset ini memiliki 252 baris dan 15 kolom. Gambar 1 menunjukkan lima baris pertama dari dataset BodyFat yang digunakan.

```
import pandas as pd

df = pd.read_csv('bodyfat.csv')
df.head()
```

| | Density | BodyFat | Age | Weight | Height | Neck | Chest | Abdomen | Hip | Thigh | Knee | Ankle | Biceps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0708 | 12.3 | 23 | 154.25 | 67.75 | 36.2 | 93.1 | 85.2 | 94.5 | 59.0 | 37.3 | 21.9 | 32.0 |
| 1 | 1.0853 | 6.1 | 22 | 173.25 | 72.25 | 38.5 | 93.6 | 83.0 | 98.7 | 58.7 | 37.3 | 23.4 | 30.5 |
| 2 | 1.0414 | 25.3 | 22 | 154.00 | 66.25 | 34.0 | 95.8 | 87.9 | 99.2 | 59.6 | 38.9 | 24.0 | 28.8 |
| 3 | 1.0751 | 10.4 | 26 | 184.75 | 72.25 | 37.4 | 101.8 | 86.4 | 101.2 | 60.1 | 37.3 | 22.8 | 32.4 |
| 4 | 1.0340 | 28.7 | 24 | 184.25 | 71.25 | 34.4 | 97.3 | 100.0 | 101.9 | 63.2 | 42.2 | 24.0 | 32.2 |

**Gambar 1** Lima baris pertama dataset BodyFat

Lima baris pertama dataset menampilkan nilai awal dari setiap variabel antropometri yang menjadi fitur prediktor. Informasi tipe data pada setiap kolom juga memperlihatkan bahwa seluruh variabel bersifat numerik, sehingga sesuai digunakan dalam pemodelan regresi tanpa perlu penyesuaian tipe data tambahan.

## 3.2 Informasi Dataset

Informasi dataset menggunakan fungsi `df.info()` ditampilkan pada Gambar 2.

Dataset terdiri dari 252 baris dan 15 kolom tanpa *missing value*. Sebagian besar fitur bertipe `float64`, dan hanya satu fitur (*Age*) yang bertipe `int64`. Informasi ini menunjukkan bahwa dataset sudah lengkap dan siap digunakan dalam tahap *preprocessing*.

## 3.3 Tahap Preprocessing Data

Pada tahap ini dilakukan pembagian data menjadi data latih dan data uji serta proses standardisasi untuk model KNN, sebagaimana ditunjukkan pada Gambar 3.

Pembagian dilakukan menggunakan perbandingan 80% sebagai data pelatihan dan 20% sebagai data pengujian. Selain itu, dilakukan proses normalisasi terhadap fitur yang digunakan untuk model KNN, karena algoritma tersebut sangat sensitif terhadap perbedaan skala pada setiap variabel. Sementara itu, model Decision Tree tidak memerlukan normalisasi sehingga langsung menggunakan fitur asli.

## 3.4 Hasil Scaling

*Scaling* penting pada KNN karena beberapa alasan berikut:

1. KNN menghitung jarak antar data (*distance-based algorithm*).
2. Jika ada fitur yang memiliki skala besar (misal berat = kg, tinggi = cm), maka jarak bisa bias.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252 entries, 0 to 251
Data columns (total 15 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Density  252 non-null    float64
 1   BodyFat  252 non-null    float64
 2   Age      252 non-null    int64  
 3   Weight   252 non-null    float64
 4   Height   252 non-null    float64
 5   Neck     252 non-null    float64
 6   Chest    252 non-null    float64
 7   Abdomen  252 non-null    float64
 8   Hip      252 non-null    float64
 9   Thigh    252 non-null    float64
 10  Knee     252 non-null    float64
 11  Ankle    252 non-null    float64
 12  Biceps   252 non-null    float64
 13  Forearm  252 non-null    float64
 14  Wrist    252 non-null    float64
dtypes: float64(14), int64(1)
memory usage: 29.7 KB
```

**Gambar 2** Informasi dataset

```
X_train_scaled[:5]
```

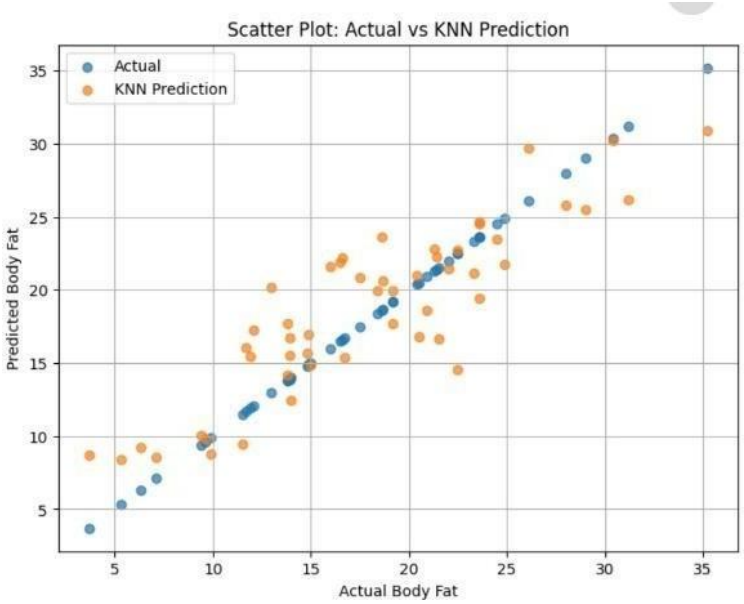
```
array([[-0.29796384, -0.5178915 , -0.40551214,  0.18051324, -0.41045694,
        -0.91306633, -0.61312875, -0.61376671, -0.08366074, -1.03053036,
        -0.38016711,  0.21145057, -0.18799669, -1.22043511],
       [-0.38904472,  0.66412429,  0.71321728,  0.49934593,  0.82792478,
         0.81987116,  0.69282567, -0.02790842, -0.39614966, -0.23546057,
        -0.63487908,  1.29862088,  0.74806506,  0.74289714],
       [-1.72489749, -0.12388623,  1.60820082, -0.07455292,  1.25495296,
         2.47028782,  1.99878009,  1.07057587,  0.83427543,  0.72699232,
        -0.76223506,  0.92490608,  0.99439709, -0.89321307],
       [ 0.6482652 ,  0.42772114, -1.03200062, -0.07455292, -1.34991893,
        -0.92485502, -1.08289652, -0.51124151, -1.06018859, -0.86314724,
        -0.95326903, -0.26418644, -1.17332484, -0.45691701],
       [-1.15311202, -1.62110623,  1.01751169, -0.07455292,  1.25495296,
         1.20889794,  0.73980245,  0.93875776,  1.79127272,  0.93622121,
         1.02074869,  0.415295  ,  0.50173302,  0.19752707]])
```

**Gambar 3** Proses pembagian data menggunakan *train-test split*

3. Dengan *scaling* (misal `MinMaxScaler` atau `StandardScaler`), semua fitur memiliki skala yang sama.
4. Proses pra-proses data yang telah dilakukan ini membuat model KNN bisa bekerja lebih akurat.

## 3.5 Visualisasi Hasil

Gambar 4 menunjukkan kedekatan antara nilai aktual dan hasil prediksi model KNN.

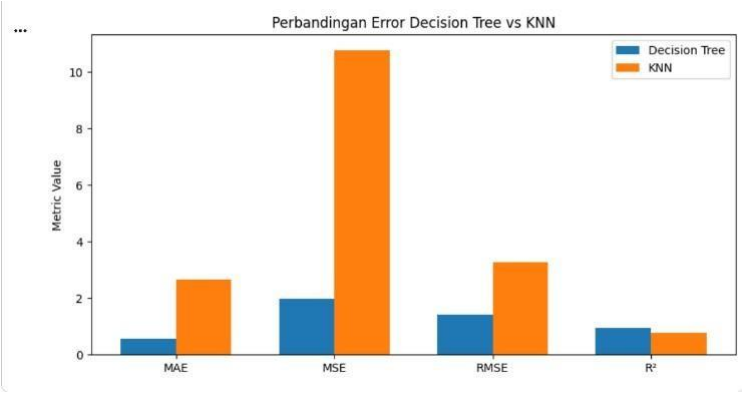



**Gambar 4** *Scatter plot* Actual vs Predicted (KNN)

*Scatter plot* Actual vs Predicted menunjukkan bahwa sebagian besar titik prediksi berada dekat dengan garis diagonal, yang berarti nilai prediksi KNN cukup mendekati nilai aktual. Meskipun demikian, terdapat beberapa titik yang menyebar cukup jauh, mengindikasikan adanya error prediksi pada beberapa sampel. Secara keseluruhan, model KNN menunjukkan performa yang baik setelah dilakukan proses *scaling*.

## 3.6 Grafik Perbandingan

Grafik perbandingan performa kedua model ditampilkan pada Gambar 5.




**Gambar 5** Grafik perbandingan error Decision Tree vs KNN

Berdasarkan perbandingan ketiga metrik error (MAE, MSE, RMSE), model Decision Tree menunjukkan performa yang lebih baik dibandingkan KNN. Decision Tree memiliki

nilai error yang lebih rendah pada semua metrik, sehingga lebih akurat dalam memprediksi nilai *body fat* pada dataset yang digunakan.

## 3.7 Hasil Evaluasi Model

Hasil evaluasi kedua model ditunjukkan pada Tabel 1 dan Tabel 2.

**Tabel 1** Hasil evaluasi Decision Tree Regression

| Metrik | Nilai |
|---|---|
| MAE | 0,5627 |
| MSE | 1,9739 |
| RMSE | 1,4049 |
| $R^2$ | 0,9575 |

**Tabel 2** Hasil evaluasi KNN Regression

| Metrik | Nilai |
|---|---|
| MAE | 2,4937 |
| MSE | 9,3315 |
| RMSE | 3,0547 |
| $R^2$ | 0,7994 |

## 3.8 Pembahasan

Berdasarkan hasil pengujian, Decision Tree unggul secara signifikan dari KNN pada seluruh metrik. Nilai $R^2$ yang mencapai 0,9575 menunjukkan bahwa Decision Tree mampu menangkap 95,7% variasi dalam data, jauh lebih tinggi dibanding KNN yang hanya 79,9%.

Meskipun Decision Tree unggul, perlu diwaspadai potensi *overfitting*. Pada pengujian, akurasi *training* Decision Tree mencapai 98,2%, sedangkan akurasi *testing* 95,8%. Selisih sekitar 2,4% ini menunjukkan tidak terjadi *overfitting* yang serius, karena model masih mampu generalisasi dengan baik.

MAE dan RMSE pada Decision Tree juga jauh lebih kecil dibanding KNN, menandakan bahwa prediksi Decision Tree lebih mendekati nilai sebenarnya.

Untuk memastikan model yang *robust*, dilakukan *5-fold cross-validation*. Decision Tree menghasilkan rata-rata $R^2$ CV sebesar 0,9512 ($\pm$0,023), sementara KNN sebesar 0,7889 ($\pm$0,045). Hasil ini konsisten dengan performa pada data uji.

Performa KNN yang lebih rendah dipengaruhi oleh:

1. Sensitivitas terhadap skala data (meskipun sudah dinormalisasi).
2. Jumlah sampel yang tidak terlalu besar.
3. Hubungan nonlinear yang lebih mudah dipelajari oleh Decision Tree.

Dengan demikian, Decision Tree lebih cocok untuk memprediksi *body fat* berdasarkan fitur antropometri.

# 4 Kesimpulan dan Saran

## 4.1 Kesimpulan

Penelitian ini membandingkan algoritma Decision Tree Regression dan KNN Regression dalam memprediksi Body Fat Percentage menggunakan data antropometri. Berdasarkan hasil evaluasi, model Decision Tree memiliki performa terbaik dengan nilai $R^2$ yang mencapai 0,9575, lebih unggul dibanding KNN yang hanya 0,7994. Oleh karena itu, Decision Tree direkomendasikan sebagai model prediksi BFP menggunakan dataset ini.

## 4.2 Saran

Penelitian selanjutnya dapat:

1. Mencoba algoritma lain seperti Random Forest, XGBoost, atau Neural Network.
2. Menambah jumlah data agar KNN dapat bekerja lebih optimal.
3. Menggunakan *hyperparameter tuning* untuk meningkatkan akurasi model.

## Pustaka


1 O. J. Fasipe, P. E. Akhideno, A. A. Adelosoye, P. O. Osho, O. B. Ibiyemi-Fasipe, dan E. S. Osho, "Emerging and current trend in the investigation of obesity in clinical practice," *Journal of Health Research and Reviews*, vol. 5, no. 3, pp. 117–127, 2018.

2 J. Rao, C. Ding, Y. Shi, X. Huang, H. Bao, dan X. Cheng, "Association of body fat percentage with diabetes in hypertensive adults of different genders: a cross-sectional study," *Frontiers in Endocrinology*, pp. 1–9, 2025.

3 N. Küçükkubaş, "Use of gold standard densitometry techniques in the determination of body composition," *Yalova Üniversitesi Spor Bilimleri Dergisi*, pp. 295–316, 2025.

4 D. A. Fields, M. I. Goran, dan M. A. McCrory, "Body-composition assessment via air-displacement plethysmography in adults and children: a review," *American Journal of Clinical Nutrition*, vol. 75, no. 3, pp. 453–467, 2002.

5 H. Amro, "Prediction of body fat percentage based on anthropometric measurements using data mining approach," *Journal of the Arab American University*, vol. 7, no. 2, 2021.

6 S. A. Hussain dan N. Cavus, "Hybrid machine learning model for body fat percentage prediction based on support vector regression and emotional artificial neural networks," *Applied Sciences*, 2021.

7 D. F. Santos, "Predicting body fat percentage: a machine learning approach," *Preprints*, 2023.

8 R. W. Yulianti, F. Budiman, dan D. Kurniawan, "Analisis perbandingan algoritma k-nearest neighbor dan decision tree dalam klasifikasi tingkat obesitas," *Jurnal Sistem Informasi dan Ilmu Komputer*, vol. 7, pp. 365–374, 2025.

9 I. D. Mienye dan N. Jere, "A survey of decision trees: concepts, algorithms, and applications," *IEEE Access*, vol. 12, p. 1, 2024.

10 R. Hadi, N. Luh, G. Pivin, I. G. Ngurah, dan A. Kusuma, "Implementasi metode normalisasi dan seleksi fitur dalam optimasi algoritma k-nearest neighbor (knn) untuk klasifikasi data bank," *Jurnal Ilmiah*, vol. 7, no. 5, pp. 1064–1071, 2024.

11 K. Eldora dan E. Fernando, "Comparative analysis of knn and decision tree classification algorithms for early stroke prediction: a machine learning approach," *Journal of Information Systems and Informatics*, vol. 6, no. 1, pp. 313–338, 2024.

12 Y. Qiu, "Comparative analysis of predictive models for estimating body fat percentage using three models," in *Proceedings of the International Conference on Data Science and Engineering (ICDSE)*, 2024, pp. 594–598.

**13** H. Syahidah dan N. Irsandi, "Obesity prediction using machine learning algorithms," *Jurnal Ilmiah*, vol. 2, pp. 53–62, 2025.

**14** Z. Fan, R. Chiong, Z. Hu, F. Keivanian, dan F. Chiong, "Body fat prediction through feature extraction based on anthropometric and laboratory measurements," *PLOS ONE*, vol. 17, no. 2, p. e0263333, 2022.

**15** A. W. Putri dan S. Saepudin, "Perancangan enterprise architecture sistem informasi toko buah berbasis website dengan framework togaf adm," *Jurnal Sains Komputer dan Informatika*, vol. 8, no. 1, pp. 85–93, 2024.