Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893

Perbandingan Metode *Decision Tree* dan *Logistic Regression* dalam Klasifikasi Tingkat Obesitas Berdasarkan Gaya Hidup

DOI: http://dx.doi.org/10.35889/jutisi.v15i2.3591

Yunisia Rosari Bere[1], Fadhli Almu'iini Ahda[2*]
Teknik Informatika, Institut Teknologi dan Bisnis Asia Malang, Malang Jawa Timur, Indonesia
**Email Corresponding Author*: fadhlial@asia.ac.id

Abstrac

This study compares two classification algorithms, *Decision Tree* and *Logistic Regression*, to predict obesity levels based on individual lifestyle patterns. The initial dataset consisted of 2,212 data points with 17 attributes, which were then narrowed down to 499 data points with the nine most relevant attributes. After a cleaning process, 498 valid data points were obtained, including demographic information and daily habits, which were then used in the modelling process. To objectively evaluate model performance, a *stratified 5-fold cross-validation* method was used, along with testing on separate test data. The evaluation results showed that *Logistic Regression* consistently performed better, with an average accuracy of 0.8755 and an F1-score of 0.8576. In contrast, *Decision Tree* achieved an accuracy of 0.7851 and an F1-score of 0.7704. The test data also showed a similar pattern, with *Decision Tree* achieving an accuracy of 0.75 and an F1-score of 0.7534, while *Logistic Regression* achieved an accuracy of 0.88 and an F1-score of 0.8664. Overall, the results showed that logistic regression performed more consistently and reliably in classifying obesity levels, suggesting that this method may be a superior method for supporting analysis in the healthcare industry.

Kata Kunci: *Classification; Obesity; Lifestyle; Decision Tree; Logistic Regression*

Abstrak

Studi ini membandingkan dua algoritma klasifikasi, *Decision Tree* dan *Logistic Regression*, untuk memprediksi tingkat obesitas berdasarkan pola gaya hidup individu. Dataset awal terdiri dari 2.212 data dengan 17 atribut, yang kemudian diseleksi menjadi 499 data dengan 9 atribut yang paling relevan. Setelah melalui proses pembersihan, diperoleh 498 data valid yang mencakup informasi demografis dan kebiasaan sehari-hari, yang selanjutnya digunakan dalam proses pemodelan. Untuk mengevaluasi kinerja model secara objektif, digunakan metode *stratified 5-fold cross-validation* serta pengujian pada data uji terpisah. Hasil evaluasi menunjukkan bahwa *Logistic Regression* secara konsisten berkinerja lebih baik, dengan rata-rata akurasi 0,8755 dan F1-score sebesar 0,8576. Sebaliknya, *Decision Tree* memperoleh akurasi 0,7851 dan F1-score sebesar 0,7704. Data pengujian juga menunjukkan pola serupa, dengan *Decision Tree* mencapai akurasi 0,75 dan F1-score sebesar 0,7534, sedangkan *Logistic Regression* mencapai akurasi 0,88 dan F1-score sebesar 0,8664. Secara keseluruhan, hasil penelitian menunjukkan bahwa regresi logistik berkinerja lebih konsisten dan andal dalam mengklasifikasikan tingkat obesitas, menunjukkan bahwa metode ini mungkin merupakan metode yang lebih unggul untuk mendukung analisis di industri perawatan kesehatan.

Kata kunci: *Klasifikasi; Obesitas; Gaya hidup; Decision Tree; Logistic Regression*

1. Pendahuluan

Karena dampaknya yang negatif dan signifikan terhadap kualitas hidup, obesitas telah menjadi masalah kesehatan utama. Penumpukan lemak tubuh yang tidak terkontrol dikaitkan dengan gangguan ini, yang seiring waktu dapat mengakibatkan sejumlah masalah kesehatan, termasuk tekanan darah tinggi, masalah jantung, dan penyakit metabolik. Organisasi Kesehatan Dunia melaporkan bahwa kasus obesitas meningkat setiap tahun dan merupakan penyebab

utama kematian dini di banyak negara [1]. Hal ini menunjukkan bahwa obesitas merupakan ancaman signifikan bagi kesehatan masyarakat, di samping juga merupakan masalah pribadi.

Banyak elemen yang saling terkait, seperti praktik kehidupan sehari-hari, aktivitas fisik, dan pola asupan makanan, berkontribusi terhadap peningkatan kejadian obesitas [2]. Menentukan tingkat obesitas sulit dilakukan karena interaksi yang rumit antara variabel-variabel tersebut. Oleh karena itu, diperlukan metode yang dapat mengidentifikasi pola yang lebih kompleks. *Machine learning* adalah metode yang dapat membantu membuat keputusan yang lebih objektif dan meningkatkan akurasi kategorisasi [3].

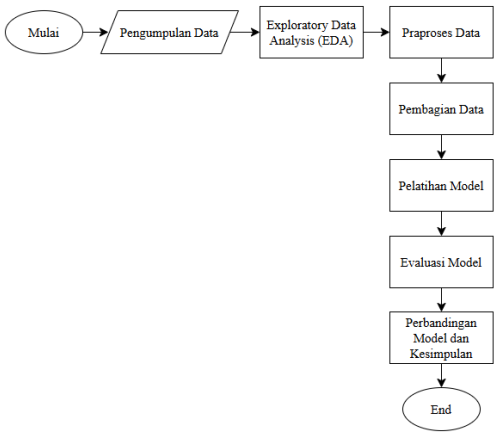
Sejumlah penelitian sebelumnya telah mengkaji tingkat obesitas menggunakan klasifikasi. Karena dapat memodelkan hubungan antar variabel, terutama pada data dengan tren linier, *Logistic Regression* sering digunakan [4], [5], [6]. Di sisi lain, *Decision Tree* sering dipilih karena dapat menghasilkan aturan keputusan yang mudah dipahami [6], [7]. Untuk mendapatkan hasil yang lebih baik, beberapa sistem alternatif juga menggabungkan teknik seperti Support Vector Machine dan K-Nearest Neighbor [3]. Namun, setiap strategi memiliki kekurangannya masing-masing. Saat berurusan dengan data yang rumit, *Decision Tree* rentan terhadap *overfitting* [7], [8], Di sisi lain, pola non-linier sulit ditangani oleh *Logistic Regression*. Selain itu, saat ini masih minim studi yang secara eksplisit membandingkan kedua pendekatan tersebut dengan penilaian yang menyeluruh [9].

Studi ini membandingkan kinerja *Decision Tree* dan *Logistic Regression* dalam mengkategorikan tingkat obesitas menggunakan metodologi evaluasi yang lebih sistematis. Untuk menciptakan model yang lebih andal dan mengurangi bias dari berbagi data, dilakukan studi yang menyeluruh [9]. Kedua pendekatan tersebut dipilih karena kualitasnya yang saling melengkapi dalam hal stabilitas model dan interpretasi [10]. Temuan studi ini dapat berfungsi sebagai panduan untuk penelitian masa depan di bidang penambangan data kesehatan dan diharapkan dapat menawarkan ringkasan teknik yang lebih efisien.

2. Metodologi

2.1 Tahap Penelitian

Untuk menghasilkan studi yang akurat, penelitian ini dilakukan dalam sejumlah langkah sistematis, sebagaimana tunjukkan pada **Gambar 1.**



Gambar 1. Alur penelitian klasifikasi tingkat obesitas berdasarkan gaya hidup

Proses penelitian dimulai dengan 499 data, kemudian dibersihkan menjadi 498 data. Selanjutnya dilakukan EDA untuk memahami pola dan distribusi data. Data dipraproses melalui encoding dan normalisasi, lalu dibagi 80:20 secara *stratified*. Model *Decision Tree dan Logistic Regression* kemudian dilatih dan dievaluasi menggunakan *5-fold cross-validation*, sebelum akhirnya dibandingkan untuk menentukan metode terbaik [4].

2.2 Dataset dan Variabel Penelitian

Data yang digunakan dalam penelitian ini berasal dari dataset publik di Kaggle yang memuat informasi terkait karakteristik individu dan gaya hidup [11]. Awalnya terdapat 2112 data dengan 17 atribut, kemudian dilakukan proses seleksi sehingga tersisa 499 data, dan setelah pembersihan diperoleh 498 data yang layak digunakan. Dengan kategori tingkat obesitas sebagai variabel target, ada 9 atribut diantaranya jenis kelamin, usia, tinggi badan, berat badan, riwayat obesitas dalam keluarga, kebiasaan konsumsi makanan berkalori tinggi, frekuensi makan utama

harian, dan tingkat aktivitas fisik, dipilih sebagai variabel yang paling relevan. Atribut yang semula berbahasa Inggris telah dialihterjemahkan ke dalam Bahasa Indonesia untuk kemudahan pemahaman. Data kemudian dibagi ke dalam data latih dan data uji dengan perbandingan 80:20 menggunakan teknik *stratified split* untuk memastikan representasi kelas tetap proporsional.

2.3 Exploratory Data Analysis (EDA)

Tahap *Exploratory Data Analysis* (EDA) bertujuan untuk mengenali pola dan karakteristik data sebelum prapemrosesan dan pemodelan, termasuk distribusi data, rentang nilai atribut, dan kondisi awal variabel target. Tahap ini penting untuk mengenali struktur data serta potensi masalah yang dapat memengaruhi kinerja model [12]. Analisis dilakukan dengan statistik deskriptif dan visualisasi histogram pada atribut seperti usia, berat badan, dan tingkat obesitas. Hasilnya menunjukkan adanya variasi nilai yang cukup besar serta ketidakseimbangan distribusi kelas, sehingga diperlukan metode evaluasi yang lebih objektif untuk setiap kelas.

2.4 Praproses Data

Setelah dataset diperoleh, dilakukan tahap pra-pemrosesan untuk menyesuaikan data agar dapat digunakan secara optimal oleh model [13]. Salah satu langkah yang dilakukan adalah mengubah variabel target tingkat obesitas dari bentuk kategorikal menjadi numerik melalui proses encoding. Dalam proses ini, kategori *Normal* dikodekan sebagai 0, *Insufficient* sebagai 1, *Overweight* sebagai 2, dan *Obesity* sebagai 3, sehingga data dapat diproses dengan lebih efektif oleh model klasifikasi [14].

Selain itu, dilakukan normalisasi untuk menyamakan skala antar fitur, mengingat Logistic Regression sensitif terhadap perbedaan skala, sedangkan *Decision Tree* tidak terlalu terpengaruh. Tahap ini bertujuan untuk mengurangi bias serta membantu meningkatkan kinerja model sebelum memasuki proses pembagian data dan pelatihan [15].

2.5 Pembagian Data

Setelah pra-pemrosesan, dataset dibagi menjadi subkelompok pelatihan dan pengujian dengan rasio 80:20, menghasilkan 398 set data pelatihan dan 100 set data pengujian. Untuk mempertahankan distribusi kelas secara proporsional di seluruh kedua subset, teknik *stratified 5-fold cross-validation* digunakan dalam prosedur ini, yang meningkatkan keterwakilan dan keandalan hasil evaluasi model yang digunakan untuk mengklasifikasikan tingkat obesitas [15].

2.6 Metode Klasifikasi

Metode klasifikasi digunakan untuk mengelompokkan tingkat obesitas berdasarkan atribut gaya hidup dan karakteristik individu. Baik *Decision Tree* maupun *Logistic Regression* digunakan dalam penelitian ini. *Decision Tree* bersifat non-parametrik dan mudah dipahami, sedangkan *Logistic Regression* lebih andal dan sering digunakan dalam klasifikasi kesehatan [16].

2.6.1 Decision Tree

Proses pengambilan keputusan direpresentasikan oleh sebuah pohon dalam metode pengkategorian yang dikenal sebagai *Decision Tree*. Setiap cabang dalam pohon mewakili proses pengurutan data berdasarkan atribut spesifik yang dianggap paling informatif. Dengan menggunakan metode ini, model dapat menghasilkan aturan keputusan yang sederhana dan membantu dalam menentukan variabel yang paling berpengaruh pada hasil kategorisasi. *Gini Index* adalah salah satu metrik yang sering digunakan untuk menilai kualitas pemisahan data selama proses pembuatan pohon keputusan. Ukuran ini menggambarkan tingkat ketidakmurnian suatu kelompok data, sehingga semakin kecil nilainya, semakin baik pemisahan yang dihasilkan pada node tersebut. Secara matematis, kriteria tersebut dirumuskan dalam suatu persamaan (1).

$$Gini(S) = 1 - \sum_{i=1}^{n} \mathrm{p}_i^2 \quad (1)$$

Keterangan:
(S) : himpunan data pada node
$\mathrm{p}i$: proporsi data pada kelas ke-i
n : jumlah kelas

Tingkat homogenitas data yang lebih tinggi ditunjukkan oleh nilai *Gini Index* yang lebih rendah, yang berkisar dari 0 hingga 1. Pemisah terbaik ditentukan dengan memilih properti dengan nilai terendah. Proses pembentukan pohon dilakukan secara berulang hingga memenuhi kondisi penghentian tertentu. Namun, model ini memiliki kelemahan berupa kecenderungan

mengalami *overfitting* apabila struktur pohon terlalu kompleks [17].

2.6.2 Logistic Regression

Salah satu teknik untuk memperkirakan kemungkinan data akan masuk ke dalam kategori tertentu adalah Logistic Regression. Model ini bekerja dengan mengolah kombinasi variabel input menjadi nilai probabilitas melalui fungsi *sigmoid*, sehingga hasil akhirnya berada pada rentang antara 0 hingga 1. Pendekatan ini cukup efektif ketika hubungan antar variabel bersifat linier [18], sebagaimana ditunjukkan pada Persamaan (2).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Nilai probabilitas dari Persamaan (2) digunakan untuk menentukan kelas data. Nilai tersebut diperoleh dari fungsi sigmoid yang dibentuk oleh kombinasi linear variabel independen pada Persamaan (3).

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3}$$

$\sigma(z)$: probabilitas data termasuk ke dalam kelas tertentu
e : bilangan Euler (≈ 2,718)
β_0 : konstanta (intercept)
β_1 ... β_n : koefisien regresi
x_1 ... x_n : variabel independen

Nilai probabilitas tersebut kemudian digunakan sebagai dasar dalam menentukan kategori kelas melalui batas ambang tertentu. Metode ini banyak diterapkan dalam bidang kesehatan karena memiliki kestabilan yang baik, mudah diinterpretasikan melalui koefisien model, serta mampu memberikan estimasi probabilitas yang jelas dalam proses klasifikasi.

2.7 Metode Evaluasi

Akurasi, presisi, recall, dan F1-score termasuk di antara metrik yang digunakan untuk menilai kinerja model. Akurasi didefinisikan sebagai proporsi prediksi yang sesuai dengan hasil aktual. Namun, karena metrik ini tidak memadai dengan data yang tidak seimbang, akurasi dan recall digunakan untuk menilai performa setiap kelas. *Confusion matrix* juga digunakan untuk memberikan ringkasan yang lebih menyeluruh tentang hasil klasifikasi. F1-score digunakan sebagai statistik utama dalam penelitian ini karena merupakan kompromi antara recall dan presisi, sehingga lebih representatif saat mengevaluasi model [19], [4], [6], [7].

2.8 Cross-validation

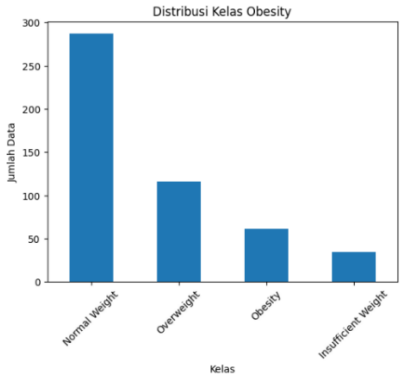
Studi ini menggunakan teknik *stratified 5-fold cross-validation* untuk mendapatkan representasi yang lebih akurat dari kinerja model. Teknik ini membagi data menjadi lima bagian menggunakan proporsi kelas yang seimbang, kemudian digunakan secara bergantian sebagai data pelatihan dan pengujian. Dengan cara ini, setiap bagian data memiliki kesempatan yang sama untuk diuji, sehingga hasil evaluasi menjadi lebih objektif. Nilai performa yang diperoleh merupakan rata-rata dari seluruh proses pengujian, sehingga mampu memberikan estimasi yang lebih konsisten dibandingkan hanya menggunakan satu kali pembagian data [20], [21], [6].

2.9 Lingkungan Pengujian

Bahasa pemrograman *Python* digunakan untuk melakukan penelitian ini. Sejumlah pustaka tambahan digunakan untuk pemrosesan dan pemodelan data, termasuk *pandas* untuk pemrosesan dan manipulasi data, *scikit-learn* untuk pembuatan dan penilaian model klasifikasi, dan *matplotlib* untuk visualisasi hasil analisis. Karena setiap eksperimen dilakukan dalam lingkungan komputasi yang umum, peneliti lain dapat dengan mudah meniru prosedur dan metode yang digunakan dalam penelitian ini.

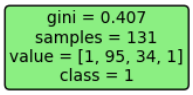
3. Hasil dan Pembahasan

Bagian ini membahas hasil pengujian model sekaligus analisis terhadap performa yang diperoleh. Dataset yang digunakan berasal dari Kaggle dengan jumlah 498 data dan 9 atribut yang merepresentasikan karakteristik gaya hidup dalam klasifikasi tingkat obesitas. Berdasarkan hasil eksplorasi data (**Gambar 2**), terlihat bahwa distribusi kelas tidak merata. Kategori *Normal Weight* menjadi yang paling dominan, kemudian diikuti oleh *Overweight*, *Obesity*, dan *Insufficient Weight*. Ketidakseimbangan ini berpotensi memengaruhi kemampuan model, terutama dalam mengenali kelas dengan jumlah data yang lebih sedikit.



Gambar 2. Distribusi Kelas Tingkat Obesitas

Pada tahap prapemrosesan, data terlebih dahulu diubah melalui proses encoding sebelum digunakan dalam pemodelan. Dalam proses evaluasi, digunakan *stratified 5-fold cross-validation* agar proporsi kelas tetap terjaga di setiap pembagian data. Penelitian ini membandingkan dua model, yaitu *Logistic Regression* dan *Decision Tree*. Decision Tree dimanfaatkan untuk menangkap pola *non-linear*, sebagaimana ditunjukkan pada visualisasi model (**Gambar 3**). Pada gambar tersebut, terdapat node dengan distribusi data yang tidak seimbang, yaitu [1, 95, 34, 1], yang didominasi oleh satu kelas tertentu. Kondisi ini membuat model cenderung mengarah pada kelas dominan, meskipun nilai *gini* sebesar 0,407 menunjukkan tingkat kemurnian yang masih tergolong cukup baik.



Gambar 3. Hasil Klasifikasi Keputusan Akhir

Decision Tree secara konsisten dikalahkan oleh *Logistic Regression*, menurut hasil evaluasi yang menggunakan *cross-validation*. *Decision Tree* hanya mencapai akurasi 0,7851 dan skor F1 0,7704, sedangkan *Logistic Regression* memperoleh akurasi rata-rata 0,8755 dengan skor F1 0,8576. Perbedaan ini menunjukkan bahwa *Logistic Regression* biasanya menunjukkan stabilitas yang lebih besar selama prosedur validasi. Kesimpulan ini juga didukung oleh hasil pengujian pada data uji yang digunakan untuk menilai kemampuan generalisasi model. *Logistic Regression* sekali lagi menghasilkan hasil yang lebih baik pada titik ini, dengan akurasi 0,88 dan skor F1 0,8664. *Decision Tree*, di sisi lain, berkinerja lebih buruk daripada *Logistic Regression* dengan akurasi 0,75 dan skor F1 0,7534.

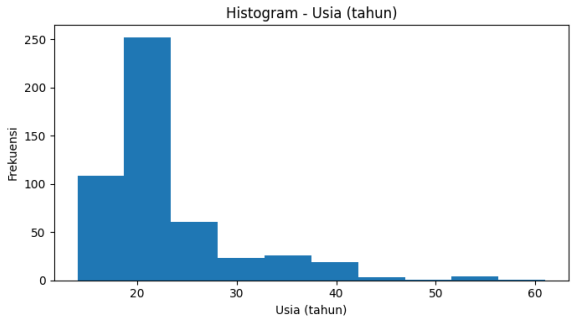
Secara analitis, kondisi ini menunjukkan bahwa data obesitas berbasis gaya hidup memiliki kecenderungan hubungan antar variabel yang relatif linier, sehingga lebih sesuai dimodelkan menggunakan *Logistic Regression*. Di sisi lain, *Decision Tree* yang mampu menangkap pola *non-linear* justru cenderung menyesuaikan diri secara berlebihan terhadap data latih. Hal ini dapat menyebabkan *overfitting* yang berdampak pada penurunan kinerja ketika model dihadapkan pada data baru. Meskipun demikian, *Decision Tree* tetap memiliki keunggulan dalam hal interpretabilitas. Bahkan bagi pengguna non-teknis seperti tenaga medis, model ini dapat menghasilkan aturan pengambilan keputusan yang lebih mudah dipahami. Selain itu, visualisasi model membuat proses pengambilan keputusan lebih mudah dipahami, sehingga pendekatan ini tetap dapat diterapkan dalam situasi yang membutuhkan transparansi. Hasil penelitian ini juga sejalan dengan beberapa studi sebelumnya yang menyatakan bahwa *Logistic Regression* lebih efektif digunakan pada data kesehatan dengan karakteristik numerik dan hubungan yang relatif linier [6], [7]. Hal ini menunjukkan bahwa pemilihan metode klasifikasi sebaiknya disesuaikan dengan karakteristik data yang digunakan.

Meskipun memberikan hasil yang cukup baik, penelitian ini masih memiliki beberapa keterbatasan, seperti ukuran dataset yang relatif kecil, ketidakseimbangan kelas, serta terbatasnya algoritma yang digunakan. Kondisi tersebut dapat memengaruhi kemampuan model dalam mengenali pola yang lebih kompleks maupun dalam mengidentifikasi kelas minoritas. Oleh karena itu, disarankan agar dilakukan studi lebih lanjut dengan menggunakan kumpulan data yang lebih besar dan seimbang, menerapkan teknik penanganan data tidak seimbang seperti SMOTE, serta mengeksplorasi metode ensemble learning seperti Random Forest atau Gradient Boosting untuk

meningkatkan performa model. Secara keseluruhan, *Logistic Regression* menunjukkan kinerja yang lebih unggul dan lebih andal dibandingkan dengan *Decision Tree*, sehingga dapat dipertimbangkan sebagai metode yang lebih sesuai untuk klasifikasi obesitas berbasis data gaya hidup. Temuan ini juga dapat menjadi acuan dalam pengembangan sistem klasifikasi di bidang kesehatan berbasis data mining [6].

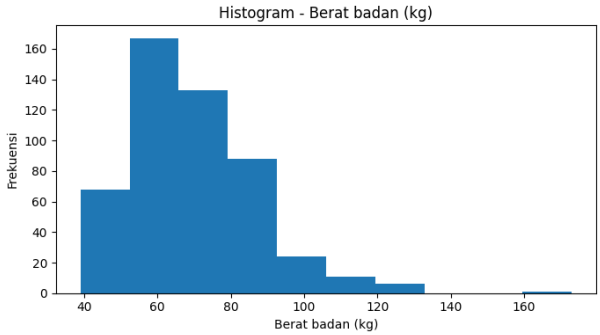
3.1 Hasil *Exploratory* Data Analysis (EDA)

EDA dilakukan untuk memahami properti asli dari dataset tersebut, khususnya distribusi variabel utama yang memengaruhi klasifikasi obesitas. Hasilnya menunjukkan bahwa usia responden didominasi kelompok dewasa muda dengan rentang 14–61 tahun, serta berat badan yang bervariasi cukup besar dari 39 kg hingga 173 kg, yang mengindikasikan perbedaan kondisi fisik antar individu.



Gambar 4. Histogram Distribusi Usia Responden

Atribut seperti jenis kelamin dan tinggi badan memiliki distribusi yang stabil tanpa nilai ekstrem. Meskipun tidak divisualisasikan, atribut tersebut tetap mendukung proses klasifikasi tingkat obesitas. Secara keseluruhan, EDA mengindikasikan adanya variasi data yang cukup, terutama pada usia dan berat badan, yang ditampilkan pada **Gambar 4** dan **Gambar 5**.



G**ambar 5.** Histrogram Distribusi Berat Badan Responden

3.2 Hasil *Cross-validation* Model

Untuk mendapatkan estimasi kinerja yang konsisten dan dapat dipercaya, teknik *stratified 5-fold cross-validation* digunakan untuk penilaian awal kinerja model. Model *Decision Tree* memiliki rata-rata akurasi 0,7851 (±0,0108) dan F1-score sebesar 0,7704 (±0,0130), menurut data, tetapi *Logistic Regression* berkinerja lebih baik dengan rata-rata akurasi 0,8755 (±0,0303) dan F1-score sebesar 0,8576 (±0,0313). Secara umum, *Logistic Regression* berkinerja lebih baik dalam mengkategorikan tingkat obesitas menurut gaya hidup.

3.3 Evaluasi Model Decision Tree pada Data Uji

Evaluasi model *Decision Tree* dilakukan menggunakan 498 data dengan 9 atribut yang dibagi menjadi 398 data latih dan 100 data uji melalui teknik *stratified split* untuk menjaga keseimbangan kelas. Hasil *stratified 5-fold cross-validation* menunjukkan akurasi rata-rata sebesar 0,7851 (±0,0108) dan F1-score sebesar 0,7704 (±0,0130). Nilai simpangan baku yang relatif kecil mengindikasikan bahwa performa model cukup stabil, meskipun masih lebih rendah dibandingkan Logistic Regression.

```
=== CV (5-Fold) - Decision Tree ===
Accuracy : 0.7851 ± 0.0108
F1-Score : 0.7704 ± 0.0130
```

Gambar 6. Hasil *Cross-validation* pada *Decision Tree*

Decision Tree mencapai akurasi 0,75, presisi 0,7646, recall 0,75, dan F1-score 0,7534 ketika model dievaluasi pada data uji untuk menilai generalisasi. Meskipun kinerjanya cukup baik, *Logistic Regression* mengungguli *Decision Tree*.

Tabel 1. Hasil Evaluasi Model *Decision Tree* pada Data Uji

Metrik	Nilai
Accuracy	0,75
Precision	0,7646
Recall	0,75
F1-Score	0,7534

Berdasarkan *classification report* pada **Gambar 7**, *Decision Tree* menunjukkan performa yang tidak seimbang, dengan kinerja baik pada kelas mayoritas (kelas 1), tetapi rendah pada kelas minoritas (kelas 0). Hal ini terlihat dari *macro F1-score* yang belum optimal, sementara *weighted F1-score* dipengaruhi oleh dominasi kelas besar.

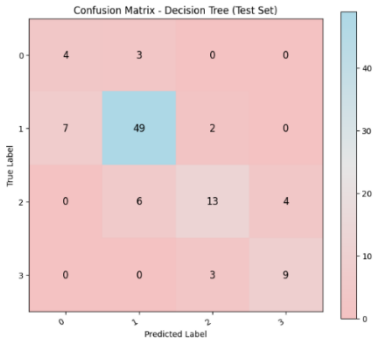
```
--- Classification Report (Decision Tree) ---
              precision    recall  f1-score   support

           0       0.36      0.57      0.44         7
           1       0.84      0.84      0.84        58
           2       0.72      0.57      0.63        23
           3       0.69      0.75      0.72        12

    accuracy                           0.75       100
   macro avg       0.66      0.68      0.66       100
weighted avg       0.76      0.75      0.75       100
```

Gambar 7. Hasil *Classification Report* pada *Decision Tree*

Confusion matrix pada **Gambar 8** menunjukkan bahwa mayoritas prediksi sudah benar, namun masih terdapat kesalahan antar kelas dengan karakteristik serupa. Hal ini menunjukkan bahwa batas pemisahan kelas belum optimal.



Gambar 8. *Confusion Matrix* Model Decision Tree pada Data Uji

Decision Tree mampu menangkap pola data, tetapi kurang baik dalam generalisasi pada data tidak seimbang sehingga performanya lebih rendah dibandingkan *Logistic Regression*. Meski demikian, model ini unggul dalam interpretabilitas karena mudah dipahami.

3.4 Evaluasi Model Logistic Regression pada Data Uji

Evaluasi *Logistic Regression* menggunakan 498 data yang dibagi secara *stratified* menjadi data latih dan uji. Hasil *5-fold cross-validation* menunjukkan akurasi 0,8755 (±0,0303) dan F1-score 0,8576 (±0,0313), yang menandakan performa model stabil dan lebih baik dibandingkan *Decision Tree*.

```
=== CV (5-Fold) - Logistic Regression ===
Accuracy : 0.8755 ± 0.0303
F1-Score : 0.8576 ± 0.0313
```

Gambar 9. Hasil *Cross-validation* pada *Logistic Regression*

Logistic Regression memperoleh akurasi 0,88, presisi 0,8907, recall 0,88, dan skor F1 0,8664 ketika model dievaluasi pada data uji untuk menilai generalisasi. Temuan ini menunjukkan bahwa model tersebut dapat menghasilkan prediksi yang seimbang dan akurat.

Tabel 2. Hasil Evaluasi Model Logistic Regression pada Data Uji

Metrik	Nilai
Accuracy	0,88
Precision	0,8907
Recall	0,88
F1-Score	0,8664

Berdasarkan *classification report* pada **Gambar 10**, model memiliki performa tinggi pada kelas dominan (kelas 1), namun masih rendah pada kelas minoritas seperti kelas 0. Meskipun demikian, nilai *weighted F1-score* yang tinggi menunjukkan bahwa performa keseluruhan tetap stabil.

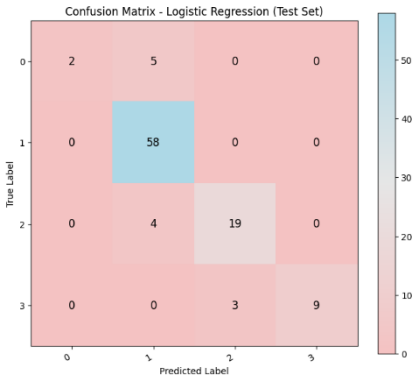
```
--- Classification Report (Logistic Regression) ---
              precision    recall  f1-score   support

           0       1.00      0.29      0.44         7
           1       0.87      1.00      0.93        58
           2       0.86      0.83      0.84        23
           3       1.00      0.75      0.86        12

    accuracy                           0.88       100
   macro avg       0.93      0.72      0.77       100
weighted avg       0.89      0.88      0.87       100
```

Gambar 10. Hasil *Classification Report* pada *Logistic Regression*

Visualisasi *confusion matrix* yang ditampilkan pada **Gambar 11** memperlihatkan bahwa jumlah prediksi yang tepat lebih dominan dibandingkan kesalahan dalam klasifikasi. Kelas dengan data yang lebih sedikit memiliki kesalahan prediksi paling banyak, yang biasanya disebabkan oleh distribusi data yang terbatas di kelas-kelas tersebut.



Gambar 11. *Confusion Matrix* Model *Logistic Regression* pada Data Uji

Secara keseluruhan, Logistic Regression menunjukkan kinerja yang baik, stabil, dan konsisten pada pelatihan maupun pengujian, serta mampu mengklasifikasikan tingkat obesitas secara andal berdasarkan faktor gaya hidup.

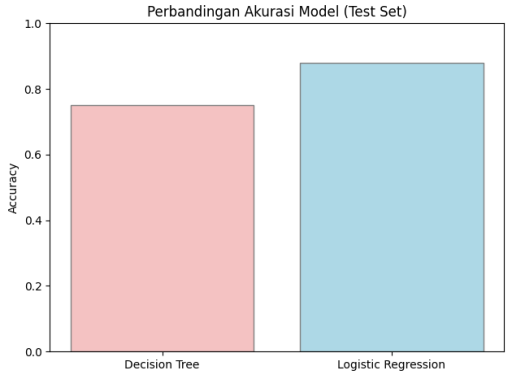
3.5 Perbandingan Kinerja Model

Berdasarkan **Tabel 3**, *Logistic Regression* menunjukkan performa yang lebih unggul dibandingkan Decision Tree baik pada *cross-validation* maupun data uji, dengan nilai akurasi dan F1-score yang lebih tinggi serta hasil yang lebih seimbang.

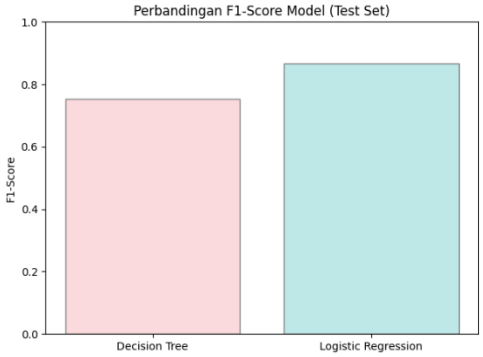
Tabel 3. Perbandingan Kinerja Model

Model	Accuracy (CV)	F1-Score (CV)	Accuracy (Test)	Precision	Recall	F1-Score (Test)
Decision Tree	0,7851 ± 0,0108	0,7704 ± 0,0130	0,7500	0,7646	0,7500	0,7534
Logistic Regression	0,8755 ± 0,0303	0,8576 ± 0,0313	0,8800	0,8907	0,8800	0,8664

Tabel 3 Hal ini menunjukkan bahwa dalam memprediksi obesitas, *Logistic Regression* lebih andal dan berhasil. Model ini stabil dari tahap validasi hingga pengujian, sehingga lebih tahan terhadap *overfitting* dibanding *Decision Tree*. Keunggulan tersebut juga terlihat pada metrik akurasi dan F1-score yang ditampilkan pada **Gambar 12** dan **Gambar 13**.



Gambar 12. Perbandingan Akurasi Model *Decision Tree* dan *Logistic Regression*



Gambar 13. Perbandingan F1-Score Model *Decision Tree* dan *Logistic Regression*

Berdasarkan analisis menyeluruh terhadap kinerja model, *Logistic Regression* adalah metode terbaik untuk memprediksi tingkat obesitas dalam dataset ini. Tingkat akurasi yang tinggi, stabilitas prediksi yang dihasilkan, dan kemampuan metode ini untuk menyelaraskan kinerja klasifikasi secara andal di semua tingkat kelas menunjukkan keunggulannya.

3.6 Pembahasan dan Hasil Penelitian

Penelitian ini memberikan kontribusi yang lebih spesifik dibandingkan studi sebelumnya dalam konteks klasifikasi tingkat obesitas. Sebagian besar penelitian terdahulu hanya berfokus pada perbandingan kinerja algoritma seperti *Logistic Regression* dan *Decision Tree* secara umum, tanpa mempertimbangkan hubungan antara karakteristik data dan kesesuaian metode yang digunakan [22], [6].

Hal ini menyebabkan masih adanya kesenjangan dalam memahami pengaruh struktur dan distribusi data, khususnya data berbasis gaya hidup, terhadap performa masing-masing algoritma klasifikasi. Keunikan studi ini terletak pada metodologi perbandingannya, yang tidak hanya mengevaluasi model menggunakan metrik penilaian, tetapi juga secara konseptual menghubungkannya dengan properti dataset. Hasil penelitian menunjukkan bahwa keunggulan

Logistic Regression tidak semata-mata berasal dari nilai akurasi yang lebih tinggi, tetapi juga karena kesesuaian asumsi model dengan pola hubungan antar variabel dalam data. Hal ini sejalan dengan literatur yang menyatakan bahwa efektivitas model sangat dipengaruhi oleh kesesuaian antara asumsi algoritma dan karakteristik data [6]. Oleh karena itu, kesesuaian model dengan struktur data dipertimbangkan ketika memilih teknik klasifikasi, selain kinerja numeriknya.

Selain itu, penelitian ini memberikan kontribusi empiris dalam klasifikasi obesitas berbasis data gaya hidup yang masih relatif terbatas dibandingkan pendekatan berbasis data klinis atau rekam medis [6]. Pendekatan ini memperluas penerapan *machine learning* dalam bidang kesehatan preventif, khususnya untuk deteksi dini risiko obesitas melalui variabel yang lebih mudah diperoleh. Penelitian ini juga menegaskan pentingnya penggunaan metrik evaluasi yang lebih representatif, yaitu F1-score, dalam menangani klasifikasi multikelas dengan data tidak seimbang, karena lebih seimbang dalam merepresentasikan *precision* dan *recall* [6], [22][23].

Secara konseptual, hasil penelitian ini memperkuat temuan sebelumnya bahwa model yang lebih sederhana seperti *Logistic Regression* dapat memberikan kinerja yang kompetitif, bahkan lebih baik dibandingkan model yang lebih kompleks, apabila sesuai dengan karakteristik data [6]. Temuan ini memiliki implikasi penting dalam pengembangan model klasifikasi di bidang *data mining* kesehatan, dengan menekankan pendekatan yang lebih adaptif dan berbasis pemahaman data. Diharapkan karya ini akan berfungsi sebagai referensi penelitian masa depan dalam analisis data kesehatan berbasis *machine learning* karena tidak hanya menawarkan bukti empiris tetapi juga memperkuat kerangka teoritis dalam memilih metode klasifikasi terbaik.

4. Simpulan

Dalam penelitian ini, *Decision Tree* dan *Logistic Regression* dibandingkan untuk klasifikasi obesitas. Hasil pengujian dan validasi menunjukkan bahwa *Logistic Regression* memberikan hasil yang lebih akurat 0,88 dan F1-score 0,8664, dibandingkan *Decision Tree* yang hanya mencapai akurasi 0,75. Keunggulan ini disebabkan kesesuaian data dengan model linier yang memberikan generalisasi lebih baik, sedangkan *Decision Tree* cenderung mengalami *overfitting*. Oleh karena itu, *Logistic Regression* lebih disarankan, sementara penelitian lebih lanjut disarankan untuk menggunakan kumpulan data yang lebih besar dan teknik ensemble untuk meningkatkan hasil.

Daftar Referensi

[1] D. Fakhrizal, A. Ananda Sutrisno, N. Afza Zain, G. Angga Mukti, and A. Setiawan, “Implementasi Algoritma Machine Learning Menggunakan Model Random Forest Untuk Klasifikasi Obesitas,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 5, pp. 7579–7584, 2025, doi: 10.36040/jati.v9i5.14667.

[2] S. Hardwis and J. Jajat, “Analisis Resiko Obesitas Berdasarkan Aktivitas Fisik: Implementasi Metode Artificial Intelligence Machine Learning,” *Jurnal Keolahragaan*, vol. 10, no. 2, p. 97, 2024, doi: 10.25157/jkor.v10i2.16884.

[3] S. A. Utiarahman, A. Mulawati, and M. Pratama, “Analisis Perbandingan KNN, SVM, Decision Tree dan Regresi Logistik Untuk Klasifikasi Obesitas Multi Kelas,” *Media Online)*, vol. 4, no. 6, pp. 3137–3146, 2024, doi: 10.30865/klik.v4i6.1871.

[4] F. Almu’iini Ahda, A. Prasetya Wibawa, D. Prasetya, and A. Sulistyo, “International Journal On Informatics Visualization Journal Homepage : Www.Joiv.Org/Index.Php/Joiv International Journal On Informatics Visualization Comparison of Adam Optimization and RMSprop in Minangkabau-Indonesian Bidirectional Translation with Neura,” vol. 8, no. March, pp. 231–238, 2024, [Online]. Available: www.joiv.org/index.php/joiv

[5] D. A. Sulistyo, D. D. Prasetya, F. A. Ahda, and A. P. Wibawa, “Pivoted Low Resource Multilingual Translation with NER Optimization,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 5, 2025, doi: 10.1145/3727876.

[6] F. A. Ahda, A. P. Wibawa, D. D. Prasetya, D. A. Sulistyo, and A. Nafalski, “Minangkabau Language Stemming: A New Approach with Modified Enhanced Confix Stripping,” *Jurnal RESTI*, vol. 9, no. 3, pp. 677–687, 2025, doi: 10.29207/resti.v9i3.6511.

[7] F. A. Ahda and M. Zainuddin, “Prediksi Kepuasan Pelayanan Perpustakaan Menggunakan Algoritma Decision Tree (C4.5),” *Jurnal Teknologi Informasi*, vol. 10, pp. 143–150, 2019, doi: 10.36382/jti-tki.v10i2.368.

[8] E. Halabaku and E. Bytyçi, “Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests,” *Intelligent Automation and Soft Computing*, vol. 39, no. 6, pp. 987–1006, 2024, doi: 10.32604/iasc.2024.059429.

[9] A. Pinar, | Fatma, H. Yagin, and | Georgian Badicu, "Use of Logistic Regression Method in Predicting Obesity Levels with Machine Learning Method," *Journal of Exercise Science & Physical Activity Reviews*, vol. 2024, no. 1, pp. 104–113, 2024, [Online]. Available: https://doi.org/10.5281/zenodo.12601115
[10] J. Pongthao, A. Na-Udom, and J. Rungrattanaubol, "Machine Learning Classification with Logistic Regression Feature Selection Approach on Health Datasets," *IAENG International Journal of Applied Mathematics*, vol. 55, no. 6, pp. 1–3, 2025.
[11] E. Dwi *et al.*, "Penggunaan Data Mining untuk Prediksi tingkat Obesitas di Meksiko Menggunakan Metode Random Forest," *Agustus*, vol. 8, pp. 2549–7952, 2024.
[12] H. W. Dhany, Sutarman, and F. Izhari, "Exploratory Data Analysis (EDA) methods for healthcare classification," *Journal of Intelligent Decision Support System (IDSS)*, vol. 6, no. 4, pp. 209–215, 2023, [Online]. Available: www.idss.iocspublisher.org
[13] P. Bangert, *MACHINE LEARNING: Konsep, Implementasi, dan Aplikasi*, vol. 45, no. 13. 2021. [Online]. Available: https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA
[14] S. Tondang, R. R. Prasetyo, R. Fulvian, Y. G. Sitorus, and G. Chrisnawati, "Analisis Perbandingan Algoritma K-Nearest Neighbor dan Ensemble Learning dalam Klasifikasi Penyakit Obesitas," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 4536–4548, 2025, doi: 10.31004/riggs.v4i2.994.
[15] R. S, B. Kumaraswamy, V. Agarwal, and A. G. Jain, "A Comparative Study of Different Data Pre-processing Methods for Machine Learning," *International Journal For Multidisciplinary Research*, vol. 7, no. 4, pp. 1–8, 2025, doi: 10.36948/ijfmr.2025.v07i04.52920.
[16] E. Z. Dahmash *et al.*, "Upholding Quality and Patient Safety during COVID-19 Pandemic—A Jordanian Case Study," *Healthcare (Switzerland)*, vol. 11, no. 4, pp. 1–13, 2023, doi: 10.3390/healthcare11040523.
[17] A. W. Wicaksono and T. Setiadi, "Penerapan Klasifikasi Decision Tree (C4.5) untuk Memprediksi Kelulusan Siswa Sekolah Dasar di Kecamatan Juai," *Format : Jurnal Ilmiah Teknik Informatika*, vol. 12, no. 2, p. 151, 2023, doi: 10.22441/format.2023.v12.i2.008.
[18] M. Fahmuddin, M. K. Aidid, and M. J. Taslim, "Implementasi Analisis Regresi Logistik Dengan Metode Machine Learning Untuk Mengklasifikasi Berita Di Indonesia," *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 5, no. 03, pp. 155–162, 2023, doi: 10.35580/variansiunm116.
[19] Y. I. Sulistya and M. Istighosah, "Obesity Prediction with Machine Learning Models Comparing Various Algorithm Performances," *International Journal of Artificial Intelligence in Medical Issues*, vol. 3, no. 1, pp. 1–13, 2025, doi: 10.56705/ijaimi.v3i1.181.
[20] G. James, T. Hastie, R. Tibshirani, and D. Witten, *An Introduction to Statistical Learning, Springer Texts*, vol. 102. 2023.
[21] H. T. Santoso, F. A. Felmidi, A. N. Fadhila, A. Ristyawan, and E. Daniati, "Analisis Kinerja Algoritma Data Mining pada Klasifikasi Tingkat Obesitas dengan K-Fold Cross Validation dan AUC," *Agustus*, vol. 8, pp. 2549–7952, 2024.
[22] M. Iwagami *et al.*, "Comparison of machine-learning and Logistic Regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study," *PLOS Digital Health*, vol. 3, no. 8, pp. 1–16, 2024, doi: 10.1371/journal.pdig.0000578.
[23] A.S. Lase, S.S. Berutu, & H. Budiati, "Implementasi Metode K-Nearest Neighbor Pada Sentimen Masyarakat Terkait Pelaksanaan KTT G20. *Progresif: Jurnal Ilmiah Komputer*, vol. 19, no. 2, pp. 481-490. 2023.