

Optimasi Algoritma Decision Tree Menggunakan GridSearchCV untuk Klasifikasi Tipe Obesitas

Feby Laurent*, Sri Winarno, Ika Novita Dewi

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: ^{1,*}112202206849@mhs.dinus.ac.id, ²sri.winarno@dsn.dinus.ac.id, ³ikadewi@dsn.dinus.ac.id

Email Penulis Korespondensi: 112202206849@mhs.dinus.ac.id

Submitted: 03/11/2025; Accepted: 09/12/2025; Published: 11/12/2025

Abstrak—Peningkatan kasus obesitas di berbagai negara, termasuk Indonesia, telah menjadi permasalahan serius dalam kesehatan masyarakat karena meningkatkan risiko penyakit kronis dan memengaruhi aspek psikologis individu. Salah satu tantangan utama dalam penanganan obesitas adalah perbedaan tipe obesitas pada setiap individu yang dipengaruhi oleh berbagai faktor, sehingga diperlukan metode klasifikasi yang akurat agar penanganan obesitas dapat lebih tepat sasaran. Dalam konteks tersebut, teknologi berbasis *machine learning* menjadi solusi potensial untuk mengklasifikasikan tipe obesitas. Namun, variasi karakteristik individu menjadikan proses klasifikasi kompleks karena model sering kesulitan membedakan tipe obesitas secara akurat. Untuk mengatasi hal tersebut, algoritma Decision Tree dipilih karena mudah diinterpretasikan hasilnya. Meskipun demikian, penggunaan Decision Tree dengan parameter default pada dataset dengan banyak atribut dan variasi tinggi cenderung menyebabkan *overfitting* serta menurunkan akurasi. Selain itu, performa Decision Tree sangat bergantung pada pengaturan *hyperparameter* sehingga perlu dilakukan teknik optimasi untuk memperoleh hasil terbaik. Berdasarkan hal tersebut, penelitian ini bertujuan mengoptimalkan algoritma Decision Tree menggunakan GridSearchCV untuk memperoleh parameter yang paling optimal sehingga dapat meningkatkan kinerja model dalam klasifikasi tipe obesitas. Dataset yang digunakan berasal dari UCI Machine Learning Repository, terdiri dari 2.111 baris data dan 17 atribut. Berdasarkan hasil pengujian awal, model default memperoleh akurasi 92,58%, recall 92,58%, precision 92,66%, dan F1-score 92,56%. Setelah dilakukan optimasi, akurasi meningkat menjadi 95,69%, recall 95,69%, precision 95,72%, dan F1-score 95,67%. Peningkatan akurasi sebesar 3,1% membuktikan efektivitas GridSearchCV dalam meningkatkan kinerja Decision Tree, menghasilkan model prediksi lebih akurat dan stabil. Penelitian ini diharapkan berkontribusi sebagai dasar pengambilan keputusan dalam deteksi dini serta pencegahan dan penanganan obesitas secara lebih efisien dan tepat sasaran.

Kata Kunci: Obesitas; Machine Learning; Klasifikasi; Decision Tree; GridSearchCV

Abstract—The rise in obesity cases in various countries, including Indonesia, has become a serious public health problem because it increases the risk of chronic diseases and affects individuals' psychological aspects. One of the main challenges in obesity management is the differences in obesity types in each individual, which are influenced by various factors. Therefore, accurate classification methods are needed to ensure more targeted treatment. In this context, machine learning-based technology is a potential solution for classifying obesity types. However, variations in individual characteristics make the classification process complex, as models often struggle to accurately distinguish obesity types. To overcome this problem, the Decision Tree algorithm was chosen because of its easy-to-interpret results. However, using Decision Tree with default parameters on datasets with many attributes and high variation tends to cause overfitting and decrease accuracy. Furthermore, Decision Tree performance is highly dependent on hyperparameter settings, requiring optimization techniques to achieve optimal results. Based on this, this study aims to optimize the Decision Tree algorithm using GridSearchCV to obtain the most optimal parameters to improve model performance in obesity type classification. The dataset used is from the UCI Machine Learning Repository, consisting of 2,111 rows of data and 17 attributes. Based on the initial test results, the default model achieved 92.58% accuracy, 92.58% recall, 92.66% precision, and 92.56% F1-score. After optimization, the accuracy increased to 95.69%, 95.69% recall, 95.72% precision, and 95.67% F1-score. The 3.1% increase in accuracy demonstrates the effectiveness of GridSearchCV in improving Decision Tree performance, resulting in a more accurate and stable prediction model. This research is expected to contribute as a basis for decision-making in early detection and prevention and treatment of obesity more efficiently and effectively.

Keywords: Obesity; Machine Learning; Classification; Decision Tree; GridSearchCV

1. PENDAHULUAN

Permasalahan kesehatan yang dihadapi masyarakat saat ini tidak hanya terbatas pada penyakit menular, melainkan juga mencakup penyakit tidak menular yang kini kian mengancam kehidupan manusia [1]. Salah satu jenis penyakit tidak menular yang menunjukkan peningkatan drastis adalah obesitas, yaitu keadaan di mana terjadi penumpukan lemak berlebih dalam tubuh akibat ketidakseimbangan antara asupan energi dan energi yang dikeluarkan [2]. Ketika asupan kalori melebihi kebutuhan energi, kelebihan tersebut akan tersimpan sebagai lemak, sehingga berat badan meningkat secara bertahap [2], [3]. Masalah ini kini menjadi isu kesehatan global yang tidak hanya dialami negara maju, tetapi juga semakin banyak terjadi di negara berkembang, termasuk Indonesia. Sejak 1975, prevalensi obesitas di dunia tercatat meningkat hampir tiga kali lipat [4]. Laporan terbaru dari *World Population Review* dan *WHO* tahun 2025 menunjukkan bahwa tingkat obesitas global terus meningkat, terutama di kawasan Pasifik dan Amerika Utara, yang menempati peringkat tertinggi dalam prevalensi obesitas, menegaskan bahwa kelebihan berat badan menjadi salah satu tantangan utama kesehatan masyarakat di tingkat global [4]. Sedangkan di Indonesia, berdasarkan laporan *World Obesity Atlas 2025*, persentase kelebihan berat badan dan obesitas meningkat cukup signifikan setiap tahunnya [5]. Pada tahun 2025, sekitar 43% populasi dewasa Indonesia diperkirakan memiliki *Body Mass Index (BMI)* tinggi, dengan 13% di antaranya tergolong obesitas. Fenomena ini dipengaruhi oleh banyak faktor seperti perilaku, genetik, kondisi psikologis, pengaruh sosial media, status ekonomi, serta karakteristik setiap individu [2], [6]. Obesitas

berdampak luas, baik secara fisik maupun psikologis. Secara fisik, kondisi ini meningkatkan risiko penyakit kronis seperti hipertensi, penyakit jantung koroner, stroke, kanker, diabetes melitus tipe 2, serta gangguan pada sistem tulang dan sendi. Sedangkan dari sisi psikologis dan sosial, obesitas dapat menurunkan kepercayaan diri, memicu stres, serta menghambat interaksi sosial [7]. Kompleksitas masalah ini semakin meningkat karena setiap individu dapat memiliki tipe obesitas yang berbeda, sehingga penanganannya memerlukan pendekatan yang lebih spesifik [8].

Dalam menghadapi permasalahan tersebut, kemampuan untuk mengklasifikasikan tipe obesitas secara akurat menjadi hal yang sangat penting karena setiap tipe obesitas memiliki tingkat risiko kesehatan yang berbeda. Melalui klasifikasi tipe obesitas, tenaga medis dan peneliti dapat mengidentifikasi kelompok individu berdasarkan tingkat keparahan obesitas serta menentukan strategi pencegahan dan penanganan yang paling sesuai [9]. Informasi ini sangat krusial dalam dunia medis karena memengaruhi jenis intervensi yang diberikan, seperti pengaturan pola makan, peningkatan aktivitas fisik, hingga pemberian terapi medis yang sesuai dengan kebutuhan individu. Selain itu, klasifikasi tipe obesitas yang tepat memungkinkan tenaga kesehatan dan peneliti untuk memantau perkembangan kondisi pasien dari waktu ke waktu, menilai efektivitas program penurunan berat badan, serta mengidentifikasi individu yang berisiko tinggi mengalami komplikasi. Oleh karena itu, penerapan metode klasifikasi tipe obesitas yang akurat diperlukan guna mendukung pengambilan keputusan berbasis data serta memastikan langkah pencegahan dan penanganan obesitas dapat dilakukan secara lebih efektif, tepat sasaran, dan berkelanjutan.

Salah satu pendekatan yang dapat mendukung hal tersebut adalah melalui penerapan algoritma *machine learning*. Algoritma *machine learning* memiliki kemampuan untuk melakukan pengolahan data berskala besar dan kompleks, sekaligus menemukan pola serta hubungan antarvariabel yang sulit terdeteksi melalui metode analisis tradisional [10], [11]. Pendekatan ini memungkinkan proses klasifikasi tipe obesitas dilakukan dengan lebih akurat dan informatif. Dalam hal ini, *Decision Tree* merupakan salah satu algoritma klasifikasi yang paling umum digunakan. Algoritma ini memiliki sifat yang sederhana, mudah diinterpretasikan, dan mampu menampilkan hasilnya dalam bentuk pohon keputusan, sehingga hubungan antarvariabel dapat dipahami dengan lebih mudah dan jelas [12], [13].

Algoritma *Decision Tree* memiliki beberapa keterbatasan yang dapat memengaruhi kinerja model. Salah satu permasalahan utama pada *Decision Tree* adalah kecenderungan model untuk mengalami *overfitting*, yaitu kondisi ketika model terlalu terfokus pada data pelatihan hingga tidak mampu mengenali pola pada data yang belum pernah dipelajari sebelumnya. Hal ini umumnya terjadi apabila struktur pohon terlalu kompleks atau terlalu dalam, sehingga model justru mempelajari *noise* pada data, alih-alih pola yang relevan. Sebaliknya, apabila parameter yang digunakan terlalu sederhana, model dapat mengalami *underfitting*, di mana model gagal menangkap kompleksitas pola data yang sesungguhnya [11]. Selain itu, hasil yang dihasilkan oleh *Decision Tree* dapat sangat bergantung pada pengaturan *hyperparameter*, seperti *max_depth*, *min_samples_split*, *min_samples_leaf*, dan *criterion*. Pemilihan kombinasi parameter yang tidak tepat dapat menyebabkan performa model menjadi kurang optimal [11], [14]. Oleh karena itu, diperlukan suatu metode optimasi yang mampu menyesuaikan kombinasi parameter *Decision Tree* secara sistematis agar diperoleh model yang lebih optimal.

Dalam penelitian-penelitian sebelumnya, berbagai algoritma *machine learning*, termasuk *Decision Tree*, telah digunakan untuk mengimplementasikan metode klasifikasi tipe obesitas menggunakan dataset yang sama dari *UCI Machine Learning Repository*. Penelitian oleh Heru Teguh Santoso (2024), mengimplementasikan beberapa algoritma seperti *Naive Bayes*, *C4.5*, *Decision Tree*, *Random Forest*, dan *K-Nearest Neighbors (KNN)* untuk membandingkan kinerja klasifikasi berdasarkan metode validasi berbeda. Akan tetapi, penelitian tersebut tidak menjelaskan proses penyesuaian (*tuning*) parameter pada model *Decision Tree* maupun algoritma lainnya [15]. Begitu pula dalam penelitian oleh Hanifatus Syahidah (2025), yang mengombinasikan beberapa algoritma, termasuk *Decision Tree*, dengan *feature selection* untuk mengurangi jumlah atribut menjadi 13 dari total 17 fitur. Meski penelitian ini menunjukkan pengaruh seleksi fitur terhadap hasil klasifikasi, proses pengaturan *hyperparameter* tidak dijelaskan, sehingga belum diketahui bagaimana konfigurasi parameter memengaruhi hasil model *Decision Tree* [16]. Selanjutnya, penelitian oleh Gregorius Airlangga (2025), membahas berbagai metode *supervised learning*, termasuk *Decision Tree*, dengan evaluasi *10-fold cross-validation*. Walaupun performa tiap algoritma dipaparkan, namun penelitian ini tidak membahas peran parameter atau strategi optimasi yang digunakan dalam meningkatkan akurasi model [17]. Penelitian lain yang dilakukan oleh Syahrazad Syaukat Al Malaky (2025), membandingkan empat algoritma yaitu *Random Forest*, *Decision Tree*, *Light Gradient Boosting Machine (LGBM)*, dan *Extra Tree* dalam klasifikasi tipe obesitas. Akan tetapi, penelitian ini tidak menjelaskan proses penentuan parameter pada masing-masing model, sehingga belum diketahui apakah performa model tersebut sudah mewakili hasil terbaik dari algoritma tersebut [18]. Terakhir, penelitian yang dilakukan oleh Dewi Nasien (2025), mengevaluasi dampak reduksi fitur menggunakan *Principal Component Analysis (PCA)* terhadap beberapa algoritma pada klasifikasi tipe obesitas. Hasil penelitian ini menunjukkan pengaruh kompleksitas fitur terhadap performa, tetapi penelitian ini tidak menjelaskan proses penentuan parameter atau upaya optimasi model yang dilakukan [19].

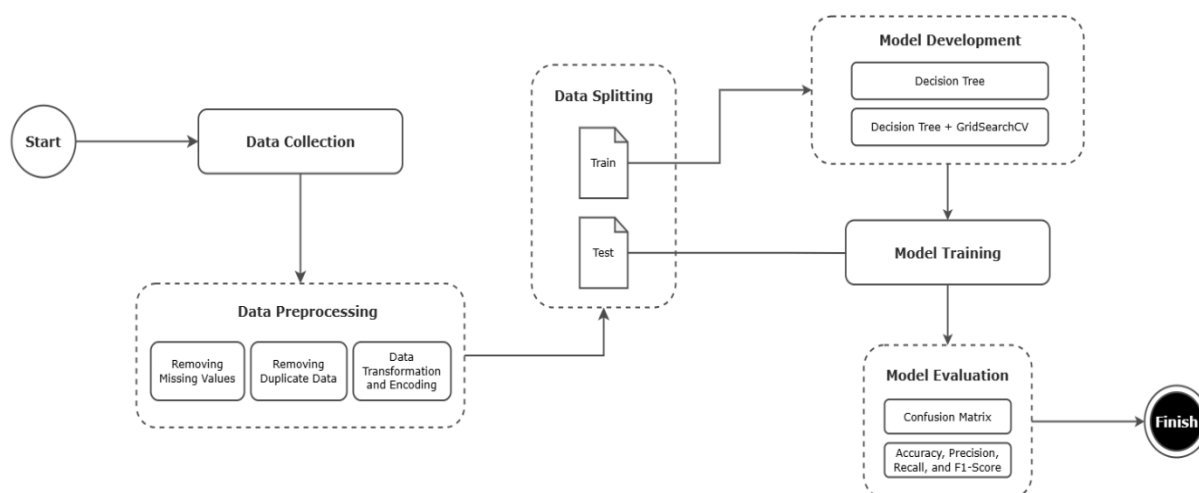
Berdasarkan berbagai penelitian sebelumnya, algoritma *Decision Tree* telah banyak dimanfaatkan dalam proses klasifikasi tipe obesitas karena kemampuannya dalam menghasilkan model yang mudah diinterpretasikan. Namun, sebagian besar penelitian tersebut hanya berfokus pada penggunaan parameter *default* tanpa menjelaskan secara rinci pengaturan *hyperparameter* yang digunakan. Selain itu, sebagian penelitian lebih menekankan pada perbandingan performa antaralgoritma tanpa mempertimbangkan penerapan teknik optimasi parameter yang sebenarnya berpotensi meningkatkan akurasi model secara signifikan. Hal ini menunjukkan bahwa aspek optimasi parameter belum menjadi fokus utama dalam penelitian-penelitian terdahulu. Kondisi ini menimbulkan celah

penelitian yang penting untuk dikaji lebih lanjut karena pemilihan parameter yang tepat memiliki pengaruh besar terhadap peningkatan performa model. Apabila proses klasifikasi tidak dilakukan dengan akurat, maka hasil prediksi dapat menyebabkan kesalahan dalam menentukan strategi penanganan yang sesuai. Kesalahan klasifikasi semacam ini dapat berdampak serius karena intervensi medis atau program pencegahan yang diterapkan mungkin tidak tepat sasaran. Oleh karena itu, peningkatan kinerja model klasifikasi menjadi aspek krusial dalam konteks kesehatan masyarakat, karena dapat membantu tenaga medis dan peneliti dalam membuat keputusan yang lebih tepat berdasarkan data yang valid dan terverifikasi. Sebagai kelanjutan dari penelitian-penelitian sebelumnya, penelitian ini bertujuan untuk menganalisis pengaruh optimasi parameter menggunakan *GridSearchCV* terhadap kinerja algoritma *Decision Tree* dalam mengklasifikasikan tipe obesitas sehingga dapat diperoleh parameter yang paling optimal. Dataset yang digunakan berasal dari *UCI Machine Learning Repository* yang terdiri dari 2.111 baris data dan 17 atribut. Temuan penelitian ini diharapkan dapat memberikan prediksi tipe obesitas dengan lebih tepat serta berkontribusi dalam pengambilan keputusan terkait pencegahan dan penanganan obesitas secara lebih efisien dan tepat sasaran.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Bagian ini menjelaskan rangkaian prosedur penelitian, dimulai dari tahap pengumpulan data hingga evaluasi model. Setiap tahap dirancang untuk menghasilkan model klasifikasi yang optimal dalam menentukan tipe obesitas berdasarkan dataset yang digunakan. Tahapan penelitian tersebut divisualisasikan melalui diagram alur penelitian sebagaimana terlihat pada Gambar 1.



Gambar 1. Alur Penelitian

Proses penelitian dimulai dengan tahap *Data Collection*, yang kemudian dilanjutkan dengan tahap *Data Preprocessing*. Pada tahap ini dilakukan proses pembersihan data dengan menghapus *missing values* dan data duplikat, serta melakukan transformasi dan *encoding*, yakni konversi variabel kategorikal menjadi format numerik agar dapat diproses lebih lanjut oleh algoritma *machine learning* [16]. Setelah itu, data dibagi menjadi dua melalui tahap *Data Splitting*, yaitu data *train* untuk pelatihan model dan data *test* untuk pengujian. Selanjutnya, pada tahap *Model Development*, dilakukan pengembangan dua model, model pertama dijalankan menggunakan *Decision Tree* dengan parameter default, kemudian dilanjutkan dengan penerapan *Decision Tree* yang dioptimasi menggunakan *GridSearchCV* guna menemukan parameter terbaik sehingga akurasi model meningkat. Kedua model tersebut kemudian melalui proses *Model Training*. Terakhir, dilakukan tahap *Model Evaluation* dengan mengukur kinerja menggunakan *Confusion Matrix*, serta metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

2.2 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berasal dari dataset publik berjudul “*Estimation of Obesity Levels Based on Eating Habits and Physical Condition*,” yang tersedia di situs *UCI Machine Learning Repository* melalui link <https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>. Jumlah keseluruhan data dalam dataset ini adalah 2.111 baris data dengan 17 atribut, yang terdiri dari 16 fitur input (*independent variables*) dan 1 fitur target (*dependent variable*) yaitu *NObesyedad* yang merepresentasikan tingkat obesitas. Atribut *NObesyedad* terbagi menjadi beberapa kategori, yaitu *Insufficient_Weight* (berat badan kurang), *Normal_Weight* (berat badan ideal), *Overweight_Level_I* dan *Overweight_Level_II* (kelebihan berat badan tingkat 1 dan 2), serta *Obesity_Type_I*, *Obesity_Type_II*, dan *Obesity_Type_III* (obesitas ringan, sedang, dan berat). Tabel 1 merupakan deskripsi spesifikasi atribut beserta jenis datanya.

Tabel 1. Atribut dan Tipe Data

Atribut	Tipe Data	Keterangan
Gender	Kategorikal	Jenis kelamin
Age	Numerik	Usia (tahun)
Height	Numerik	Tinggi badan (m)
Weight	Numerik	Berat badan (kg)
family_history_with_overweight	Biner	Riwayat obesitas keluarga
FAVC	Biner	Asupan makanan tinggi kalori
FCVC	Numerik	Intensitas konsumsi sayur
NCP	Numerik	Frekuensi makan harian
CAEC	Kategorikal	Konsumsi camilan
SMOKE	Biner	Kebiasaan merokok
CH2O	Numerik	Asupan air per hari (L)
SCC	Biner	Pemantauan kalori harian
FAF	Numerik	Aktivitas fisik per minggu (jam)
TUE	Numerik	Durasi penggunaan gadget per hari
CALC	Kategorikal	Konsumsi alkohol
MTRANS	Kategorikal	Mode transportasi utama
NObesydad	Kategorikal	Tingkat obesitas

2.3 Data Preprocessing

Proses *data preprocessing* diterapkan untuk memastikan dataset yang digunakan telah bersih, konsisten, dan siap digunakan dalam proses pemodelan. Tujuan utama tahap ini adalah untuk meminimalkan kesalahan yang berpotensi memengaruhi hasil analisis sekaligus meningkatkan kualitas data, sehingga model yang dikembangkan mampu mencapai tingkat akurasi yang lebih baik. *Preprocessing* data dilakukan melalui beberapa tahapan berikut:

a. Penanganan *Missing Values*

Pada tahap penanganan *missing value*, dilakukan pemeriksaan untuk mendeteksi nilai kosong pada dataset. Nilai yang hilang diisi menggunakan rata-rata (*mean*) untuk data numerik dan modus untuk data kategorikal. Jika jumlah nilai kosong terlalu banyak dan berpotensi mengganggu hasil analisis, maka fitur tersebut dihapus dari dataset.

b. Penghapusan Data Duplikat

Langkah berikutnya yaitu dilakukan pembersihan data duplikat agar tidak terdapat baris data sama yang muncul lebih dari satu kali agar hasil pelatihan model tidak bias dan tetap akurat. Setelah melalui tahap ini, dataset dianggap bersih sehingga siap untuk tahap transformasi dan pemodelan data berikutnya.

c. Transformasi dan *Encoding Data*

Tahap transformasi dan *encoding* data berfungsi untuk mengubah seluruh fitur dalam dataset menjadi bentuk numerik, agar siap diproses oleh algoritma *machine learning*. Pada tahap ini, fitur-fitur kategorikal seperti Gender, CALC, FAVC, SCC, SMOKE, family_history_with_overweight, CAEC, MTRANS, dan NObesydad dikonversi menjadi nilai numerik menggunakan metode *label encoding*. Teknik ini dipilih karena bersifat sederhana dan efisien, terutama untuk fitur yang memiliki dua atau beberapa kategori tanpa urutan tertentu. Sebagai contoh, kategori “Male” dan “Female” masing-masing diubah menjadi angka 0 dan 1 tanpa mengubah maknanya. Sebelum proses *encoding* dilakukan, setiap fitur kategorikal terlebih dahulu diidentifikasi dengan cermat untuk memastikan konversi berjalan dengan benar dan hasil transformasi sesuai dengan struktur data yang dibutuhkan dalam pemodelan [16].

2.4 Data Splitting

Tahap *data splitting* dilakukan dengan membagi dataset menjadi dua bagian, yaitu *data training* dan *data testing*. Langkah ini bertujuan agar model dapat mempelajari pola dari data pelatihan serta menguji kemampuan prediksinya pada data yang belum pernah digunakan sebelumnya. Proses tersebut penting untuk menilai kemampuan model dalam mengenali pola baru serta mencegah terjadinya *overfitting* pada hasil pelatihan [20]. *Data training* dimanfaatkan untuk membangun model dan menyesuaikan parameter berdasarkan pola yang ditemukan dalam dataset. Sementara itu, *data testing* digunakan untuk menilai tingkat kemampuan model dalam memprediksi data baru yang tidak digunakan selama proses pelatihan. Pada penelitian ini, dataset dibagi dengan rasio 80% untuk *training* dan 20% untuk *testing*, karena perbandingan tersebut dianggap mampu menjaga keseimbangan antara proses pembelajaran dan pengujian model sesuai dengan jumlah data yang ada. Melalui tahap ini, diperoleh gambaran yang lebih objektif mengenai performa model dalam melakukan klasifikasi tipe obesitas.

2.5 Algoritma Decision Tree

Decision Tree merupakan metode klasifikasi dalam *machine learning* yang membangun model prediksi berupa struktur pohon hierarkis. Setiap simpul pada pohon mewakili pengujian atribut tertentu, sementara setiap cabang menunjukkan hasil keputusan hingga mencapai simpul daun yang berisi label kelas [12]. Sebagai bagian dari algoritma



machine learning, *Decision Tree* bekerja dengan pendekatan alur bercabang seperti *flowchart*, di mana setiap cabang atau simpul merepresentasikan aturan keputusan tertentu (*decision rules*). Aturan-aturan ini digunakan untuk memandu proses klasifikasi berdasarkan data yang diberikan. Namun kualitas hasil klasifikasi sangat bergantung pada kelengkapan dan keakuratan data yang digunakan dalam proses pelatihan. Oleh karena itu, syarat utama dalam penerapan metode *Decision Tree* adalah tersedianya ruang data atau sampel yang akan diklasifikasikan, lengkap dengan deskripsi atribut, nilai, serta tujuan klasifikasinya [12]. Algoritma *Decision Tree* memanfaatkan konsep *Entropy* dan *Information Gain* untuk memilih atribut terbaik dalam memisahkan data. *Entropy* digunakan untuk mengukur tingkat ketidakpastian atau variasi data pada suatu node. Nilainya tinggi jika data memiliki banyak variasi kelas, dan nilainya nol jika data homogen. Rumus *Entropy* dapat dilihat pada rumus (1).

$$Entropy(S) = \sum_{i=0}^n - p_i * \log^2 p_i \tag{1}$$

Pada persamaan (1) di atas, S mewakili himpunan data yang dianalisis, n adalah jumlah kelas yang terdapat dalam dataset, dan pi menunjukkan probabilitas dari kelas ke-i, yang diperoleh dari perbandingan jumlah data pada kelas tersebut terhadap total data di S. Setelah nilai *entropy* dihitung, tahap berikutnya adalah menentukan *Information Gain*, yang berfungsi untuk menilai seberapa besar pengurangan ketidakpastian setelah data dibagi berdasarkan atribut tertentu. Rumus *Information Gain* dapat dilihat pada rumus (2).

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \tag{2}$$

Pada rumus (2) diatas, S adalah dataset awal, A adalah atribut yang diuji, Si merupakan subset data yang terbentuk setelah pembagian berdasarkan nilai atribut A, |Si| menunjukkan jumlah data dalam subset tersebut, dan |S| adalah total jumlah data dalam dataset awal. Nilai *entropy* dari masing-masing subset dikalikan dengan proporsi jumlah datanya untuk mendapatkan *entropy* gabungan setelah pemisahan. Nilai *entropy* gabungan kemudian dikurangi dari *entropy* awal untuk memperoleh *Information Gain*. Atribut dengan *Information Gain* tertinggi dipilih untuk membentuk cabang pada pohon keputusan [21].

2.6 GridSearchCV

GridSearchCV merupakan salah satu fitur pada modul *Scikit-Learn* yang berfungsi untuk melakukan pencarian kombinasi parameter terbaik (*hyperparameter tuning*) pada suatu model secara sistematis dan otomatis. Proses ini diawali dengan menentukan beberapa parameter yang akan diuji beserta rentang nilai yang mungkin menghasilkan performa terbaik. Selanjutnya, *GridSearchCV* membentuk sebuah “*grid*” yang berisi seluruh kemungkinan kombinasi dari nilai parameter-parameter tersebut. Setiap kombinasi kemudian diuji melalui proses *k-fold cross-validation*, di mana data pelatihan dibagi menjadi k bagian (*fold*). Pada setiap iterasi, satu *fold* berperan sebagai data validasi, sedangkan bagian lainnya digunakan untuk melatih model. Model *Decision Tree* kemudian dilatih dan diuji pada setiap kombinasi parameter di seluruh *fold* untuk memperoleh rata-rata nilai performa, seperti akurasi [22]. Dengan cara ini, *GridSearchCV* mampu menilai secara menyeluruh bagaimana setiap kombinasi parameter memengaruhi hasil prediksi model. Setelah semua kombinasi diuji, sistem akan memilih konfigurasi dengan nilai akurasi rata-rata tertinggi sebagai parameter terbaik yang dianggap paling optimal untuk model tersebut. Dalam penelitian ini, parameter yang diuji pada algoritma *Decision Tree* melalui *GridSearchCV* ditampilkan pada Tabel 2. Kombinasi parameter terbaik hasil dari proses *GridSearchCV* inilah yang kemudian digunakan untuk membangun model akhir *Decision Tree*, sehingga diharapkan dapat menghasilkan klasifikasi tipe obesitas yang lebih akurat dan andal.

Tabel 2. Parameter Setting

Parameter	Deskripsi	Nilai yang Diuji (GridSearchCV)
criteria	Fungsi pengukuran kualitas pemisahan	["gini", "entropy"]
max_depth	Batas maksimum tingkat kedalaman pohon	[None, 5, 10, 15, 20]
min_samples_split	Jumlah minimum data untuk pemisahan node	[2, 5, 10]
min_samples_leaf	Jumlah minimum data pada daun keputusan	[1, 2, 4]
cv	Jumlah fold untuk cross-validation	5
scoring	Metrik evaluasi model	"accuracy"

2.7 Model Evaluation

Evaluasi kinerja model dilakukan menggunakan confusion matrix, yaitu matriks yang mengukur kemampuan model klasifikasi dengan membandingkan hasil prediksi terhadap nilai sebenarnya pada data pengujian [11]. *Confusion matrix* ditampilkan pada Tabel 3.

Tabel 3. Confusion Matrix

Predicted Class	Actual Class	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)



Confusion matrix terdiri dari empat komponen utama, yaitu *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*. *True Positive* menunjukkan jumlah data yang benar-benar positif dan diprediksi sebagai positif oleh model, sementara *True Negative* menggambarkan data yang sesungguhnya negatif dan diklasifikasikan dengan tepat sebagai negatif. *False Positive* muncul saat model secara keliru mengenali data yang seharusnya negatif sebagai data positif (sering disebut *false alarm*), sementara *False Negative* menggambarkan kondisi ketika data positif justru diklasifikasikan negatif (dikenal sebagai *miss*) [11]. Berdasarkan hasil *confusion matrix*, beberapa metrik evaluasi seperti akurasi, *precision*, *recall*, dan *F1-score* dapat dihitung [23]. Rumus perhitungan akurasi, *precision*, *recall*, dan *F1-score* disajikan pada persamaan berikut.

a. Akurasi

Akurasi menunjukkan tingkat ketepatan model dalam mengklasifikasikan seluruh data uji. Nilai ini menunjukkan proporsi data yang diklasifikasikan dengan benar dibandingkan dengan keseluruhan data yang ada. Semakin tinggi akurasi, semakin baik kemampuan model dalam mengenali pola dan memberikan prediksi yang tepat.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

b. Precision

Precision mengukur tingkat ketepatan hasil prediksi positif yang dihasilkan model terhadap kondisi aktual. Tingginya nilai *precision* menandakan model jarang salah menandai data negatif sebagai positif.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

c. Recall

Recall menunjukkan sejauh mana model mampu mendeteksi seluruh data yang benar-benar termasuk dalam kelas positif. Nilai *recall* yang tinggi menunjukkan kemampuan model dalam mengenali seluruh data yang relevan, meskipun mungkin terjadi sedikit kesalahan pada klasifikasi kelas lain.

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

d. F1-Score

F1-score adalah metrik yang mengombinasikan *precision* dan *recall* melalui rata-rata harmonik. Nilai *F1-score* yang tinggi mencerminkan keseimbangan performa model yang optimal, baik dalam mendeteksi data positif maupun menjaga ketepatan prediksi.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

3. HASIL DAN PEMBAHASAN

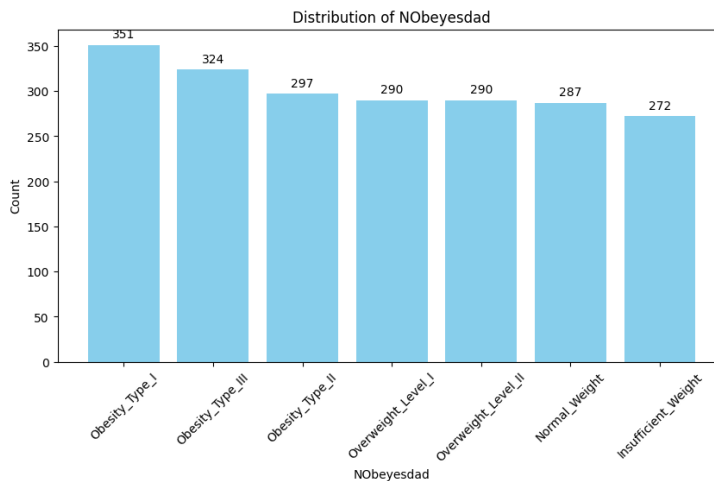
3.1 Pengumpulan Data

Proses pengumpulan data dilakukan dengan cara mengunduh dataset dalam format CSV melalui situs resmi *UCI Machine Learning Repository* yang berisi 2.111 baris data dengan 17 atribut. Selanjutnya, data tersebut diimpor dan diolah menggunakan bahasa pemrograman *Python* dengan bantuan pustaka *Pandas* untuk menjalankan proses pengolahan serta analisis lebih lanjut. Informasi dataset ditampilkan pada Tabel 4 berikut.

Tabel 4. Informasi Dataset

No	Age	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH2O	family_histo ry_with_ove rweight	FAF	TUE	CAEC	MTRANS	NOBeyesdad
0	21	Female	1.62	64	no	no	2	3	no	no	2	yes	0	1	Someti mes	Public_Tra nsportation	Normal_Wei ght
1	21	Female	1.52	56	Someti mes	no	3	3	yes	yes	3	yes	3	0	Someti mes	Public_Tra nsportation	Normal_Wei ght
2	23	Male	1.8	77	Freque ntly	no	2	3	no	no	2	yes	2	1	Someti mes	Public_Tra nsportation	Normal_Wei ght
3	27	Male	1.8	87	Freque ntly	no	3	3	no	no	2	no	2	0	Someti mes	Walking	Overweight_ Level_I
4	22	Male	1.78	89.8	Someti mes	no	2	1	no	no	2	no	0	0	Someti mes	Public_Tra nsportation	Overweight_ Level_II
...
2106	20.9 8	Female	1.71	131.41	Someti mes	yes	3	3	no	no	1.73	yes	1.68	0.91	Someti mes	Public_Tra nsportation	Obesity_Type _III
2107	21.9 8	Female	1.75	133.74	Someti mes	yes	3	3	no	no	2.01	yes	1.34	0.6	Someti mes	Public_Tra nsportation	Obesity_Type _III
2108	22.5 2	Female	1.75	133.69	Someti mes	yes	3	3	no	no	2.05	yes	1.41	0.65	Someti mes	Public_Tra nsportation	Obesity_Type _III
2109	24.3 6	Female	1.74	133.35	Someti mes	yes	3	3	no	no	2.85	yes	1.14	0.59	Someti mes	Public_Tra nsportation	Obesity_Type _III
2110	23.6 6	Female	1.74	133.47	Someti mes	yes	3	3	no	no	2.86	yes	1.03	0.71	Someti mes	Public_Tra nsportation	Obesity_Type _III

Selanjutnya, dilakukan analisis terhadap distribusi data berdasarkan variabel target, yaitu NObesyesdad. Analisis ini bertujuan untuk melihat proporsi jumlah data pada setiap kategori tingkat obesitas sehingga dapat diketahui apakah dataset memiliki distribusi yang seimbang atau tidak. Distribusi yang seimbang penting untuk memastikan bahwa model klasifikasi tidak bias terhadap kelas tertentu dan mampu melakukan pembelajaran dengan performa yang optimal. Visualisasi distribusi data pada setiap kategori tingkat obesitas ditunjukkan pada Gambar 2.



Gambar 2. Distribusi Variabel Target

Berdasarkan diagram diatas, kategori *Obesity_Type_I* (obesitas tipe ringan) memiliki jumlah data terbanyak, yaitu sebanyak 351 data, diikuti oleh *Obesity_Type_III* (obesitas berat) dengan 324 data, serta *Obesity_Type_II* (obesitas tipe sedang) sebanyak 297 data. Sementara itu, kategori *Overweight_Level_I* (kelebihan berat badan tingkat 1) dan *Overweight_Level_II* (kelebihan berat badan tingkat 2) memiliki jumlah data yang hampir sama, masing-masing 290 data, disusul oleh *Normal_Weight* (berat badan ideal) dengan 287 data, dan kategori *Insufficient_Weight* (berat badan kurang, di bawah berat normal) dengan jumlah paling sedikit yaitu 272 data. Distribusi ini menunjukkan bahwa persebaran data pada setiap kelas relatif seimbang, karena perbedaan jumlah antar kelas tidak terlalu signifikan. Kondisi ini mengindikasikan bahwa dataset memiliki proporsi yang baik antar kategori, sehingga dapat mendukung proses pelatihan model klasifikasi secara optimal tanpa risiko bias terhadap kelas tertentu.

Selanjutnya, dilakukan analisis deskriptif statistik untuk memberikan pemahaman umum mengenai distribusi dan karakteristik data, baik pada variabel numerik maupun kategorikal. Analisis ini dilakukan untuk memastikan bahwa data yang digunakan memiliki distribusi yang wajar, dapat merepresentasikan kondisi sebenarnya, tidak terdapat nilai ekstrem yang mengganggu, serta untuk memahami pola awal sebelum dilakukan proses pemodelan lebih lanjut. Hasil eksplorasi data ditampilkan pada Gambar 3 dan Gambar 4.

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111	2111	2111	2111	2111	2111	2111	2111
mean	24.31	1.7	86.59	2.42	2.69	2.01	1.01	0.66
std	6.35	0.09	26.19	0.53	0.78	0.61	0.85	0.61
min	14	1.45	39	1	1	1	0	0
25%	19.95	1.63	65.47	2	2.66	1.58	0.12	0
50%	22.78	1.7	83	2.39	3	2	1	0.63
75%	26	1.77	107.43	3	3	2.48	1.67	1
max	61	1.98	173	3	4	3	3	2

Gambar 3. Statistik Deskriptif untuk Variabel Numerik

	Gender	CALC	FAVC	SCC	SMOKE	family_history_with_overweight	CAEC	MTRANS	NObesyesdad
count	2111	2111	2111	2111	2111	2111	2111	2111	2111
unique	2	4	2	2	2	2	4	5	7
top	Male	Sometimes	yes	no	no	yes	Sometimes	Public_Transportation	Obesity_Type_I
freq	1068	1401	1866	2015	2067	1726	1765	1580	351

Gambar 4. Statistik Deskriptif untuk Variabel Kategorikal

Hasil analisis menunjukkan bahwa data numerik memiliki sebaran yang relatif merata tanpa adanya nilai ekstrem yang mencolok. Begitu pula pada variabel kategorikal, distribusi antar kategori menunjukkan keseimbangan

proporsi yang cukup baik. Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa dataset yang digunakan memiliki karakteristik data yang bersih dan representatif, sehingga layak untuk dijadikan dasar dalam proses analisis dan pemodelan pada tahap berikutnya.

3.2 Data Preprocessing

Tahap *data preprocessing* dilakukan untuk memastikan dataset telah bersih dan siap dipakai pada tahap pemodelan. Tahapan ini meliputi pemeriksaan terhadap nilai yang hilang (*missing values*) serta data duplikat. Berdasarkan evaluasi, seluruh atribut dalam dataset bebas dari nilai kosong, sehingga tidak diperlukan proses imputasi data. Selanjutnya, dilakukan identifikasi terhadap baris data yang terduplikasi untuk menghindari bias pada proses pembelajaran model. Ditemukan sebanyak 24 baris data duplikat dari total 2.111 baris data awal, kemudian baris duplikat tersebut dihapus sehingga jumlah data akhir menjadi 2.087 baris. Dari hasil tersebut dapat disimpulkan bahwa dataset telah bersih, tidak mengandung nilai kosong maupun duplikasi data, sehingga siap digunakan pada tahap transformasi dan pemodelan berikutnya. Perbandingan kondisi dataset sebelum dan sesudah proses pembersihan ditampilkan pada Tabel 5.

Tabel 5. Hasil Preprocessing

Keterangan	Sebelum Pembersihan	Setelah Pembersihan
Jumlah Data	2111	2087
Missing Values	0	0
Data Duplikat	24	0

Setelah tahap pembersihan data selesai, dataset siap untuk melalui proses transformasi dan *encoding* agar seluruh fitur dapat digunakan oleh algoritma *machine learning*. Transformasi ini terutama ditujukan pada variabel kategorikal, yaitu Gender, CALC, FAVC, SCC, SMOKE, *family_history_with_overweight*, CAEC, MTRANS, dan NObeyesdad. Pada proses ini, setiap kategori dalam atribut dikonversi menjadi representasi numerik menggunakan metode *Label Encoding*. Metode ini dipilih karena sederhana, efisien, dan sesuai untuk variabel yang memiliki beberapa kategori tanpa urutan tertentu. Transformasi ini memastikan data tetap representatif, sekaligus memenuhi format numerik yang dibutuhkan oleh algoritma klasifikasi. Gambar 5 menampilkan contoh sebagian data sebelum proses transformasi dan encoding.

	Age	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH20	family_history_with_overweight	FAF	TUE	CAEC	MTRANS	NObeyesdad	
0	21	0	1.62	64	0	0	2	3	0	0	2		1	0	1	1	4	1
1	21	0	1.52	56	1	0	3	3	1	1	3		1	3	0	1	4	1
2	23	1	1.8	77	2	0	2	3	0	0	2		1	2	1	1	4	1
3	27	1	1.8	87	2	0	3	3	0	0	2		0	2	0	1	0	2
4	22	1	1.78	89.8	1	0	2	1	0	0	2		0	0	0	1	4	3

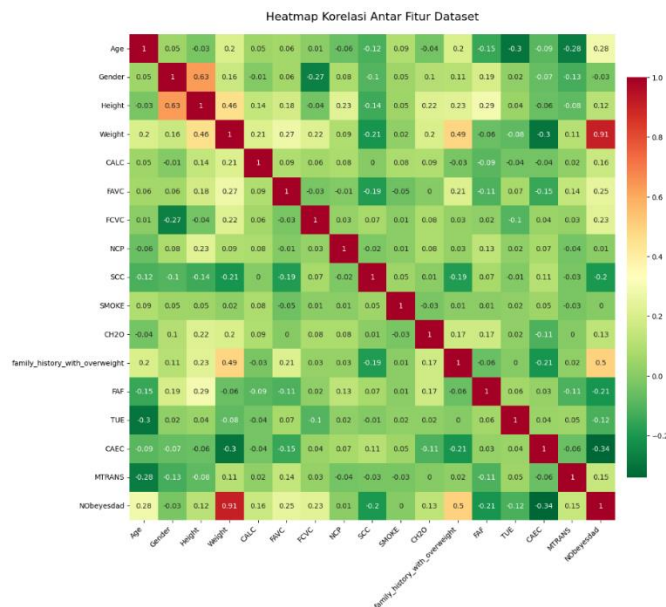
Gambar 5. Data Sebelum Transformasi dan Encoding

Setelah dilakukan label encoding, semua variabel kategorikal diubah menjadi bilangan numerik, sehingga dataset dapat langsung diproses oleh algoritma *machine learning*. Gambar 6 menunjukkan hasil transformasi dan *encoding*. Dari Gambar 6, terlihat bahwa setiap kategori berhasil dikonversi menjadi angka tanpa mengubah makna aslinya. Dataset yang telah melalui proses transformasi dan *encoding* ini telah siap untuk tahap pemodelan dan analisis klasifikasi obesitas, dengan format numerik yang sesuai dan representatif.

	Age	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH20	family_history_with_overweight	FAF	TUE	CAEC	MTRANS	NObeyesdad	
0	21	0	2	64	0	0	2	3	0	0	2		1	0	1	1	4	1
1	21	0	2	56	1	0	3	3	1	1	3		1	3	0	1	4	1
2	23	1	2	77	2	0	2	3	0	0	2		1	2	1	1	4	1
3	27	1	2	87	2	0	3	3	0	0	2		0	2	0	1	0	2
4	22	1	2	90	1	0	2	1	0	0	2		0	0	0	1	4	3

Gambar 6. Data Setelah Transformasi dan Encoding

Setelah seluruh fitur dikonversi ke dalam bentuk numerik, dilakukan analisis korelasi antar variabel untuk mengidentifikasi hubungan antar atribut dalam dataset. Analisis ini berguna untuk mengidentifikasi keterkaitan antar atribut yang dapat memengaruhi proses klasifikasi. Hubungan antar variabel divisualisasikan melalui heatmap korelasi yang disajikan pada Gambar 7.



Gambar 7. Encoding Heatmap Korelasi Antar Fitur

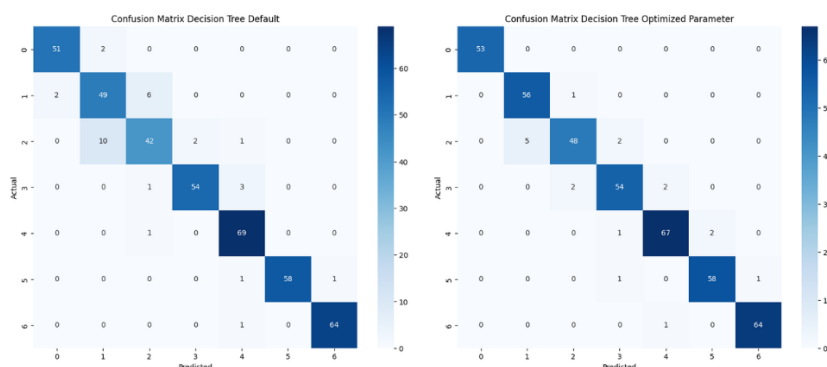
Dari Gambar 7, terlihat bahwa sebagian besar fitur memiliki korelasi yang cenderung rendah, sehingga masing-masing variabel dapat dianggap memiliki kontribusi yang relatif mandiri terhadap model. Namun, terdapat beberapa atribut yang menunjukkan korelasi cukup tinggi, seperti antara Weight dan NOBeyesdad sebesar 0,91. Hal ini mengindikasikan bahwa berat badan memiliki keterkaitan yang kuat dengan tingkat obesitas serta tinggi badan responden. Sementara itu, fitur lain seperti Age, Gender, dan Height menunjukkan korelasi yang rendah terhadap variabel target, sehingga pengaruhnya terhadap hasil klasifikasi dianggap minimal. Secara keseluruhan, temuan ini mengindikasikan bahwa dataset memiliki hubungan antar fitur yang seimbang dan dapat digunakan pada tahap pemodelan berikutnya.

3.3 Evaluasi Hasil

Proses pemodelan dalam penelitian ini dilaksanakan dengan menerapkan algoritma *Decision Tree* pada dataset yang telah melewati tahap persiapan sebelumnya. Evaluasi dilakukan dengan membandingkan kinerja model *Decision Tree* yang menggunakan parameter *default* dengan model yang telah disesuaikan melalui proses optimasi menggunakan *GridSearchCV*. Perbandingan ini bertujuan untuk mengetahui sejauh mana penyesuaian parameter dapat meningkatkan performa dan akurasi model dalam melakukan klasifikasi.

Pada model dengan parameter *default*, tingkat akurasi pada data uji mencapai 92,58%. Berdasarkan *confusion matrix* untuk model *Decision Tree* dengan parameter *default* yang ditampilkan pada Gambar 8, terlihat bahwa mayoritas data telah diklasifikasikan secara benar, meskipun masih terdapat kesalahan pada beberapa kelas, terutama kelas dengan jumlah data lebih kecil seperti kelas 1 dan kelas 2. Temuan tersebut mengindikasikan bahwa model *default* masih belum mampu secara optimal membedakan karakteristik pada setiap kelas.

Langkah berikutnya adalah melakukan optimasi parameter menggunakan *GridSearchCV* untuk memperoleh konfigurasi terbaik pada algoritma *Decision Tree*. Hasil optimasi menunjukkan peningkatan akurasi pada data uji menjadi 95,69%. Selain itu, hasil *confusion matrix* pada Gambar 8 menunjukkan bahwa jumlah prediksi benar pada setiap kelas meningkat dan kesalahan klasifikasi berkurang secara signifikan. Hasil ini menunjukkan bahwa proses optimasi berhasil meningkatkan akurasi model serta kemampuan dalam mengenali pola data secara lebih tepat.



Gambar 8. Hasil Confusion Matrix

Adapun parameter terbaik yang diperoleh dari hasil optimasi menggunakan *GridSearchCV* ditampilkan pada Tabel 6 dibawah ini, yang memuat konfigurasi optimal untuk meningkatkan kinerja model *Decision Tree*.

Tabel 6. Parameter terbaik

Parameter	Nilai Optimal
criteria	entropy
max_depth	None
min_samples_leaf	2
min_samples_split	2

Evaluasi model kemudian dilanjutkan secara menyeluruh dengan memanfaatkan metrik akurasi, *recall*, *precision*, dan *F1-score* yang dihitung dari *confusion matrix*. Hasil classification report untuk setiap kelas ditampilkan dalam bentuk visualisasi pada Gambar 9, sedangkan perbandingan nilai metrik evaluasi disajikan pada Tabel 7.

Classification Report Decision Tree Default:					Classification Report Decision Tree Optimized::				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.9623	0.9623	0.9623	53.0000	0	1.0000	1.0000	1.0000	53.0000
1	0.8033	0.8596	0.8305	57.0000	1	0.9180	0.9825	0.9492	57.0000
2	0.8400	0.7636	0.8000	55.0000	2	0.9412	0.8727	0.9057	55.0000
3	0.9643	0.9310	0.9474	58.0000	3	0.9310	0.9310	0.9310	58.0000
4	0.9200	0.9857	0.9517	70.0000	4	0.9571	0.9571	0.9571	70.0000
5	1.0000	0.9667	0.9831	60.0000	5	0.9667	0.9667	0.9667	60.0000
6	0.9846	0.9846	0.9846	65.0000	6	0.9846	0.9846	0.9846	65.0000
accuracy	0.9258	0.9258	0.9258		accuracy	0.9569	0.9569	0.9569	
macro avg	0.9249	0.9219	0.9228	418.0000	macro avg	0.9570	0.9564	0.9563	418.0000
weighted avg	0.9266	0.9258	0.9256	418.0000	weighted avg	0.9572	0.9569	0.9567	418.0000

Gambar 9. Classification Report

Tabel 7. Metrik Evaluasi

Model	Akurasi	Recall	Precision	F1-Score
Decision Tree Default	92,58%	92,58%	92,66%	92,56%
Decision Tree Optimized (GridSearchCV)	95,69%	95,69%	95,72%	95,67%

Berdasarkan metrik evaluasi yang ditampilkan pada Tabel 6, model *Decision Tree* dengan parameter *default* memperoleh nilai akurasi sebesar 92,58%, *recall* 92,58%, *precision* 92,66%, dan *F1-score* 92,56%. Hasil ini mengindikasikan bahwa model *default* telah mampu melakukan klasifikasi dengan tingkat akurasi yang tergolong tinggi. Namun, setelah dilakukan optimasi *hyperparameter* menggunakan *GridSearchCV*, seluruh metrik evaluasi mengalami peningkatan dengan akurasi sebesar 95,69%, *recall* 95,69%, *precision* 95,72%, dan *F1-score* 95,67%, yang berarti terjadi peningkatan kinerja model sebesar kurang lebih 3,1% dibandingkan dengan model default. Peningkatan nilai pada semua metrik tersebut membuktikan bahwa proses *hyperparameter tuning* menggunakan *GridSearchCV* memberikan dampak positif terhadap performa model. Dari temuan ini dapat disimpulkan bahwa proses optimasi parameter berperan penting dalam meningkatkan efektivitas algoritma *Decision Tree* dalam melakukan klasifikasi tipe obesitas secara lebih akurat dan andal. Hasil ini memiliki implikasi penting dalam konteks permasalahan obesitas, karena model klasifikasi yang lebih akurat dapat membantu mengidentifikasi tipe obesitas tiap individu dengan lebih tepat. Dengan demikian, strategi pencegahan dan intervensi medis dapat disesuaikan secara lebih spesifik berdasarkan karakteristik individu. Model yang lebih andal juga mendukung pengambilan keputusan berbasis data dalam bidang kesehatan, sehingga upaya penanganan obesitas dapat dilakukan secara lebih efektif, efisien, dan tepat sasaran.

3.4 Perbandingan dengan Penelitian Sebelumnya

Bagian ini membahas hasil perbandingan penelitian dengan sejumlah studi sebelumnya yang juga menggunakan algoritma klasifikasi untuk mendeteksi tipe obesitas. Perbandingan ini bertujuan untuk menilai sejauh mana penerapan optimasi parameter menggunakan *GridSearchCV* pada algoritma *Decision Tree* mampu meningkatkan performa model dibandingkan penelitian terdahulu, yang umumnya masih menggunakan pengaturan parameter bawaan atau belum menjalankan proses tuning secara menyeluruh. Hasil perbandingan tersebut disajikan pada Tabel 8.

Tabel 8. Perbandingan dengan Penelitian Terdahulu

No.	Peneliti	Metode	Optimasi Model	Performa Terbaik			
				Akurasi	Recall	Precision	F1-Score
1.	[15]	NB, KNN, DT, RF, C4.5	-	95%	-	-	-
2.	[16]	KNN, NBC, SVM, RF, DT	-	90,54%	88.83%	90.74%	-

No.	Peneliti	Metode	Optimasi Model	Performa Terbaik			
				Akurasi	Recall	Precision	F1-Score
3.	[17]	LR, RF, GB, AB, SVM, DT, KNN, NV	-	93,6%	93,8%	93,5%	93,5%
4.	[18]	RF, LGBM, DT, ET	-	92%	92%	92%	-
5.	[19]	KNN, SVM, DT, RF, GB, XGB	-	89%	92%	87%	89%
6.	Penelitian ini	DT Default	-	92,58%	92,58%	92,66%	92,56%
		DT Optimized	GridSearchCV	95,69%	95,69%	95,72%	95,67%

Berdasarkan perbandingan hasil algoritma *Decision Tree* yang ditampilkan pada tabel di atas, dapat diketahui bahwa penelitian ini yang menerapkan *Decision Tree Optimized* menggunakan *GridSearchCV* berhasil mencapai nilai akurasi sebesar 95,69%, *recall* 95,69%, *precision* 95,72%, dan *F1-score* 95,67%, yang merupakan hasil tertinggi dibandingkan dengan penelitian-penelitian sebelumnya. Sebagian besar studi terdahulu hanya memperoleh akurasi pada kisaran 89% hingga 95%, dan beberapa di antaranya tidak mencantumkan secara lengkap nilai metrik evaluasi seperti *precision*, *recall*, maupun *F1-score*. Temuan ini menunjukkan bahwa proses optimasi *hyperparameter* dengan metode *GridSearchCV* memberikan kontribusi nyata terhadap peningkatan kinerja model, baik dari segi ketepatan klasifikasi maupun keseimbangan antar metrik evaluasi. Dengan hasil tersebut, dapat disimpulkan bahwa pendekatan yang diterapkan dalam penelitian ini lebih unggul dalam menghasilkan model klasifikasi tipe obesitas yang tidak hanya memiliki tingkat akurasi yang tinggi, tetapi juga menunjukkan kestabilan dan konsistensi yang lebih baik dibandingkan dengan metode pada penelitian sebelumnya.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma *Decision Tree* yang dioptimasi menggunakan *GridSearchCV* terbukti mampu meningkatkan kinerja model dalam melakukan klasifikasi tipe obesitas secara signifikan. Proses optimasi *hyperparameter* memberikan dampak positif terhadap hasil evaluasi, di mana seluruh metrik evaluasi meliputi akurasi, *recall*, *precision*, dan *F1-score* mengalami peningkatan dibandingkan model dengan parameter *default*. Model *Decision Tree default* menghasilkan nilai akurasi sebesar 92,58%, *recall* 92,58%, *precision* 92,66%, dan *F1-score* 92,56%, sedangkan model yang telah dioptimasi dengan *GridSearchCV* mencapai akurasi 95,69%, *recall* 95,69%, *precision* 95,72%, dan *F1-score* 95,67%. Kenaikan sebesar sekitar 3,1% ini menunjukkan bahwa *GridSearchCV* efektif dalam menemukan kombinasi parameter terbaik untuk meningkatkan ketepatan serta konsistensi model dalam melakukan prediksi. Selain itu, penelitian ini juga menegaskan bahwa kualitas data serta tahapan *preprocessing* yang tepat, seperti penghapusan data duplikat dan transformasi fitur kategorikal menjadi numerik, berperan penting dalam mendukung keberhasilan proses klasifikasi. Hasil penelitian ini memperkuat bukti bahwa algoritma *Decision Tree* masih menjadi metode yang efektif dalam klasifikasi tipe obesitas, khususnya ketika disertai dengan penerapan optimasi parameter yang sistematis. Temuan ini diharapkan dapat memberikan sistem prediksi tipe obesitas yang lebih akurat, sehingga dapat dimanfaatkan sebagai alat bantu dalam pengambilan keputusan di bidang kesehatan, terutama dalam upaya pencegahan dan penanganan obesitas secara lebih efisien, cepat, dan tepat sasaran. Sebagai saran untuk penelitian selanjutnya, disarankan untuk menggabungkan algoritma *Decision Tree* dengan metode optimasi lain seperti *RandomizedSearchCV* guna memperoleh hasil yang lebih komprehensif.

REFERENCES

- [1] P. S. Nugroho, "Jenis Kelamin Dan Umur Berisiko Terhadap Obesitas Pada Remaja Di Indonesia," *An-Nadaa: Jurnal Kesehatan Masyarakat*, vol. 7, no. 2, p. 110, 2020, doi: 10.31602/ann.v7i2.3581.
- [2] L. Dewi and Rr. A. Ayuningtyas, *A-Z Tentang Obesitas*. UGM PRESS, 2023.
- [3] S. K. Saraswati *et al.*, "Literature Review : Faktor Risiko lingkungan," *Media Kesehatan Masyarakat Indonesia*, vol. 20, no. 1, pp. 70–74, 2021.
- [4] Nadia Putri, "10 Negara dengan Tingkat Obesitas Tertinggi di Dunia, Apakah Indonesia Termasuk," *mediaindonesia.com*. Accessed: Oct. 12, 2025. [Online]. Available: <https://mediaindonesia.com/humaniora/781284/10-negara-dengan-tingkat-obesitas-tertinggi-di-dunia-apakah-indonesia-termasuk>
- [5] World Obesity Atlas, "World Obesity Atlas 2025 | World Obesity Federation," *worldobesity.org*. Accessed: Oct. 12, 2025. [Online]. Available: <https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2025>
- [6] W. Kurdanti *et al.*, "Jurnal gizi klinik Indonesia.," *Jurnal Gizi Klinik Indonesia*, vol. 15, no. 1, pp. 22–27, 2015, [Online]. Available: <https://journal.ugm.ac.id/jgki/article/view/34939/24736>
- [7] D. S. Kohir, A. Murhan, and S. Sulastri, "Skrining Faktor Risiko Obesitas Usia Produktif," *Jurnal Wacana Kesehatan*, vol. 9, no. 2, p. 97, 2024, doi: 10.52822/jwk.v9i2.673.
- [8] A. B. Yamantri and A. A. Rifa'i, "Penerapan Algoritma C4.5 Untuk Prediksi Faktor Risiko Obesitas Pada Penduduk Dewasa," *Jurnal Komputer Antartika*, vol. 2, no. 3, pp. 118–125, 2024, doi: 10.70052/jka.v2i3.341.



- [9] M. Yhogha Ismail Ibn Ibrahim and Sandi Badiwibowo Atim, “Klasifikasi Level Obesitas Menggunakan Decision Tree C45 Dalam Menentukan Akurasi Pada Kriteria Information Gain, Gain Ratio, Gini Index,” *JIKI (Jurnal Ilmu Komputer & Informatika)*, vol. 5, no. 2, pp. 169–178, 2024, doi: 10.24127/jiki.v5i2.7191.
- [10] S. A. Utiahman and A. M. M. Pratama, “Analisis Perbandingan KNN, SVM, Decision Tree dan Regresi Logistik Untuk Klasifikasi Obesitas Multi Kelas,” *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 6, pp. 3137–3146, 2024, doi: 10.30865/klik.v4i6.1871.
- [11] C. Ihsan, *Dasar-Dasar Machine Learning Teori, Algoritma, dan Implementasi*. Greenbook Publisher, 2025.
- [12] D. Jollyta, Prihandoko, A. Hajjah, E. Haerani, and M. Siddik, *Algoritma Klasifikasi untuk Pemula Solusi Python dan RapidMiner*. Deepublish, 2023.
- [13] M. G. Febrian, R. Ferdiansyah, E. A. Nugraha, D. Satriatama, and R. Kusumastuti, “Prediksi Risiko Diabetes Menggunakan Algoritma Decision Tree Dengan Aplikasi Rapid Miner,” *Jurnal Teknologi Informasi*, vol. 2, no. November, pp. 14–24, 2024.
- [14] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, “Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5,” *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, pp. 64–71, 2020.
- [15] 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.
- [16] H. Syahidah, N. Irsandi, A. N. Ajizah, and Amelia, “Obesity Prediction Using Machine Learning Algorithms,” *Indonesian Journal of Applied Technology and Innovation Science*, vol. 2, no. February, pp. 53–62, 2025, doi: 10.1201/9781032725444-13.
- [17] G. Airlangga, “Machine Learning-Based Obesity Classification: A Comparative Study Using Self-Reported Survey Data and Ensemble Learning Models,” *Jurnal Teknologi Informatika dan Komputer*, vol. 11, no. 1, pp. 347–361, 2025, doi: 10.37012/jtik.v11i1.2585.
- [18] S. S. Al Malaky, A. A. C. Nisa, S. Armiyanti, and R. S. Setyawan, “Comparative Study of Obesity Levels Classification,” *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 10, no. 1, pp. 69–75, 2025, doi: 10.54732/jeeecs.v10i1.8.
- [19] D. Nasien *et al.*, “Optimization of Body Mass Index Classification Using Machine Learning Approach for Early Detection of Obesity Risk,” *Journal of Applied Business and Technology*, vol. 6, no. 3, pp. 193–200, 2025, doi: 10.35145/jabt.v6i3.201.
- [20] J. T. M. A. Nazanah and M. I. Jambak, “Pemanfaatan Algoritma Decision Tree ID3 Bagi Manajemen Bimbel Untuk Menentukan Faktor Kelulusan Pada Sekolah Kedinasan,” *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 6, pp. 915–924, 2023, doi: 10.30865/klik.v3i6.791.
- [21] I. Arfyanti, M. Fahmi, and P. Adytia, “Penerapan Algoritma Decision Tree Untuk Penentuan Pola Penerima Beasiswa KIP Kuliah,” *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, pp. 1196–1201, 2022, doi: 10.47065/bits.v4i3.2275.
- [22] Z. Maisat Eka Darmawan and A. Fauzan Dianta, “Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM,” *Teknologi*, vol. 13, no. 1, pp. 8–15, 2023, doi: 10.26594/teknologi.v13i1.3098.
- [23] A. Yaqin, D. Kurniawan, and J. Zeniarja, “Optimasi Algoritma K-Nearest Neighbors Menggunakan GridSearchCV untuk Klasifikasi Penyakit Diabetes,” *Infotekmesin*, vol. 16, no. 1, pp. 75–84, 2025, doi: 10.35970/infotekmesin.v16i1.2557.