JURNAL KHATULISTIWA INFORMATIKA
p-ISSN: 2339-1928 e-ISSN: 2579-633X

VOL. 12 NO. 1 Desember 2024
Halaman : 83 – 91

Klasifikasi Obesitas dengan Algoritma C5.0 Berdasarkan Pola Makan dan Kondisi Fisik

Muhammad Nasim [1]; Alda Cendekia Siregar [2]; Rachmat Wahid Saleh Insani [3]

Prodi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer
Universitas Muhammadiyah Pontianak
201220036@unmuhpnk.ac.id[1] , alda.siregar@unmuhpnk.ac.id[2], rachmat.wahid@unmuhpnk.ac.id[3]

INFO ARTIKEL

Diajukan :
05 September 2024

Diterima :
10 September 2024

Diterbitkan:
19 Desember 2024

Kata Kunci :
Algoritma C5.0 , Akurasi, Confusio Matrix, Obesitas, Klasifikasi.

INTISARI

Obesitas merupakan salah satu masalah kesehatan global yang terus meningkat, dipengaruhi oleh berbagai faktor seperti gaya hidup yang tidak sehat, pola makan tinggi kalori, dan kurangnya aktivitas fisik. kondisi ini dapat menyebabkan berbagai komplikasi serius seperti penyakit jantung, diabetes tipe 2, tekanan darah tinggi, dan berbagai kondisi kesehatan lainnya yang mengurangi kualitas hidup dan meningkatkan angka kematian. dalam penelitian ini, kami mengembangkan sistem klasifikasi tingkat obesitas menggunakan algoritma c5.0, yang dikenal karena kemampuannya dalam menangani data yang kompleks dan multikategori. algoritma ini juga efektif dalam menghasilkan model pohon keputusan yang mudah diinterpretasi oleh tenaga kesehatan. dataset yang digunakan dalam penelitian ini terdiri dari 2.111 sampel dengan 17 variabel, termasuk jenis kelamin, usia, tinggi badan, berat badan, kebiasaan makan, riwayat keluarga, dan aktivitas fisik. model c5.0 yang dibangun menunjukkan hasil yang sangat baik dengan akurasi mencapai 94,78% pada data uji. evaluasi model dilakukan menggunakan matriks kebingungan yang menunjukkan performa tinggi dengan nilai akurasi, presisi, recall, dan f1-score yang konsisten di hampir semua kategori obesitas. secara khusus, model ini mencapai nilai sempurna dalam mendeteksi kategori obesity type iii, yang menunjukkan kemampuannya yang kuat dalam mengidentifikasi tingkat obesitas yang paling parah. hasil ini menunjukkan bahwa algoritma c5.0 dapat menjadi alat yang efektif untuk mendukung sistem pendukung keputusan dalam mendeteksi risiko obesitas, yang pada akhirnya dapat membantu dalam pengembangan strategi pencegahan dan intervensi yang lebih efektif untuk meningkatkan kesehatan masyarakat.

I. PENDAHULUAN

Obesitas terjadi akibat asupan kalori yang berlebihan, yang menyebabkan penumpukan lemak tubuh dengan cepat. Kondisi ini muncul ketika jumlah kalori yang dikonsumsi melebihi jumlah kalori yang digunakan oleh tubuh(Alpiansah & Ramdhani, 2023). Obesitas kini menjadi isu kesehatan serius di seluruh dunia, dengan lebih dari 1,9 miliar orang dewasa yang mengalami kelebihan berat badan, dan sekitar 600 juta di antaranya menderita obesitas, menurut Organisasi Kesehatan Dunia (WHO). Laporan dari Survei Kesehatan dan Morbiditas Nasional juga menunjukkan bahwa perempuan lebih rentan mengalami obesitas dibandingkan laki-laki, dengan angka prevalensi masing-masing 29,6% untuk perempuan dan 25% untuk laki-laki(Susindra & Permatasari, 2023). Penyebab obesitas meliputi berbagai faktor, termasuk biologis, perkembangan, lingkungan, perilaku, serta faktor genetik. Komorbiditas yang sering terjadi pada orang dewasa dengan obesitas antara lain diabetes melitus tipe 2, hipertensi, penyakit hati berlemak non-alkohol, apnea tidur obstruktif, dislipidemia, dan sindrom metabolik. Kondisi ini juga meningkat seiring dengan meningkatnya prevalensi obesitas pada anak-anak dan remaja. Selain itu, obesitas pada anak-anak dan remaja dapat menyebabkan gangguan psikologis seperti depresi, kecemasan, rendah diri, masalah sosial, dan gangguan makan. Ketidak seimbangan energi yakni asupan kalori yang lebih banyak dibandingkan kalori yang dikeluarkan merupakan penyebab utama obesitas pada anak-anak dan remaja(Alexander Halim Santoso et al., 2023).Karena itu, menjadi penting untuk mengembangkan metode yang efisien dalam memprediksi dan mencegah obesitas. Data mining merupakan teknik yang berguna untuk menganalisis kumpulan data besar dan rumit guna mengenali pola dan anomali. Salah satu alat yang

sering digunakan dalam klasifikasi adalah algoritma *C5.0*(Wijaya et al., 2018).

Beberapa penelitian terkait algoritma C5.0 menunjukkan hasil yang bervariasi dalam penerapannya, seperti studi Rizkya yang menggunakan data 108 mahasiswa dan menghasilkan akurasi 91% untuk klasifikasi data UKT(Amalda et al., 2022). Studi lain oleh Devi membandingkan algoritma C5.0 dengan SVM dan Naïve Bayes dalam prediksi banjir, dengan hasil akurasi SVM dan C5.0 mencapai 93,75%, lebih tinggi dibandingkan Naïve Bayes yang hanya mencapai 81,25%(Fitrianah et al., 2022). Penelitian yang dilakukan oleh Natanael Benediktus menjelaskan tentang performa akademik siswa, data yang digunakan diambil dari LMS yang memiliki variable numerik dan kategorik. Algoritma C5.0 digunakan pada penelitian ini yang menghasilkan nilai akurasi sebesar 71,667% dengan data training sebesar 75% dan data testing 25%(Benediktus & Oetama, 2020). Menurut Nurnia Zamasi, dalam jurnal yang berjudul "Implementasi Algoritma C 5.0 Pada Analisa Data Potensi Pertanian dan Perternakan" menunjukkan bahwa, penerapan algoritma C5.0 dapat memberi gambaran potensi pertanian dan peternakan di setiap wilayahnya dengan menggunakan aplikasi *RapidMiner Classification Decision Tree*(Zamasi, 2021).

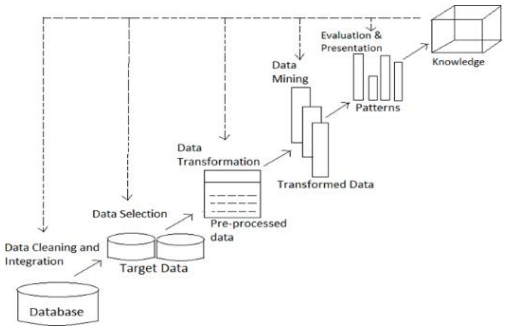
Algoritma C5.0 memiliki kemampuan untuk mengklasifikasikan sebagai pohon keputusan atau seperangkat aturan. Kekuatan algoritma ini dalam menangani nilai, serta lebih sedikit waktu diperlukan untuk mempelajarinya. C5.0 adalah *classifier* yang mengklasifikasikan data dalam waktu yang lebih singkat dibandingkan dengan *classifier* lain. Untuk menghasilkan pohon keputusan penggunaan memori minimum dan juga meningkatkan akurasi. C5.0 merupakan terusan dari algoritma C4.5, dimana C5.0 memperbaiki kelemahan dari C4.5 dari segi hasil klasifikasi, kesalahan klasifikasi, prediksi akurasi, waktu, tingkat kesalahan, dan pemakaian memori(Utomo et al., 2020).

Tujuan penelitian ini adalah untuk memanfaatkan Teknik data mining, terutama algoritma C5.0 , untuk melakukan klasifikasi tingkat obesitas dengan menggunakan *Bahasa pemograman R*. Data yang digunakan dalam penelitian berasal dari https://www.kaggle.com/datasets/ankurbajaj9/obesity levels dataset. Dataset ini mencakup informasi tentang individu dari Kolombia, Peru, dan Meksiko, termasuk faktor risiko obesitas serta tingkat obesitas yang mereka alami.

Penelitian ini bertujuan untuk memberikan kontribusi signifikan dalam penanggulangan masalah obesitas melalui pengembangan model klasifikasi yang akurat. Dengan model ini, individu yang rentan terhadap obesitas dapat diidentifikasi dan intervensi pencegahan yang tepat dapat diterapkan. Artikel ini menyajikan hasil penelitian tentang penggunaan algoritma C5.0 untuk klasifikasi obesitas, dengan data yang diambil dari sumber ini dan memanfaatkan fungsi-fungsi yang tersedia di *R*.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Databases (KDD)*, yang merupakan metode komprehensif yang terdiri dari berbagai tahapan untuk menggali, menganalisis, dan mengeksplorasi data yang besar dan kompleks. *KDD* adalah proses yang sistematis dan terstruktur yang bertujuan untuk mengekstraksi data berharga yang sebelumnya tidak terdeteksi atau tersembunyi dalam kumpulan data yang besar. Tujuan utama dari metode ini adalah untuk mengidentifikasi pola, tren, atau pengetahuan baru yang dapat memberikan wawasan berguna dan memiliki potensi untuk diterapkan dalam berbagai bidang (Firmansyah & Nurdiawan, 2023). Berikut ini adalah langkah-langkah yang akan diambil dalam penerapan metode *KDD:*



Sumber: Hasil penelitian (2024)

Gambar 1 Metode KDD

Metode *KDD* Gambar 1 menunjukkan tahapan dalam *Knowledge Discovery in Databases (KDD),* yang meliputi Pemilihan Data, Pra-pemrosesan, Transformasi, Penambangan Data, serta Interpretasi dan Evaluasi.

2.1. Data Selection

Pemilihan Data Langkah pertama adalah pemilihan data yang relevan untuk penelitian ini. Data yang digunakan dalam penelitian ini berasal dari dataset yang tersedia di https://www.kaggle.com/datasets/ankurbajaj9/obesity. Dataset ini mencakup informasi individu dari Meksiko, Peru, dan Kolombia, termasuk faktor risiko obesitas serta tingkat obesitas mereka.

2.2. Preprocessing
Pra-pemrosesan Pada tahap pra-pemrosesan, dilakukan penanganan terhadap data yang memiliki nilai-nilai hilang *(missing values)* serta seleksi atribut yang penting. Langkah ini bertujuan untuk meningkatkan kualitas data dengan mengisi atau menghapus nilai-nilai yang hilang dan memilih atribut yang paling relevan dan signifikan. Pembersihan dan persiapan data dalam penelitian ini mencakup penamaan ulang kolom. Selain itu, data dinormalisasi untuk memastikan bahwa semua fitur berada dalam skala yang sama.

2.3. Transformation
Pada tahap ini, dilakukan transformasi data yang sangat penting untuk mengubah data ke dalam format yang sesuai untuk proses eksplorasi dan analisis, yang dikenal sebagai data mining. Tujuan dari proses ini adalah untuk menyesuaikan struktur dan karakteristik data agar dapat digunakan secara optimal dalam algoritma dan teknik analisis yang diterapkan dalam data mining. Selama proses ini, dataset yang telah disiapkan dibagi menjadi dua bagian, yaitu data latih *(training data)* dan data uji *(testing data)*, dengan perbandingan 7:3. Setelah itu, algoritma C5.0 diterapkan pada dataset yang memuat informasi tentang obesitas. Algoritma C5.0 adalah algoritma klasifikasi yang digunakan dalam data mining, khususnya dalam teknik pohon keputusan *(decision tree).* Algoritma ini merupakan pengembangan dari algoritma sebelumnya, yaitu ID3 dan C4.5, yang dikembangkan oleh Ross Quinlan pada tahun 1987. Proses pembentukan pohon keputusan pada algoritma C5.0 mirip dengan algoritma C4.5, terutama dalam perhitungan *entropy* dan *information gain.* Namun, terdapat perbedaan signifikan pada langkah lanjutan setelah perhitungan *information gain.* Pada algoritma C4.5, perhitungan berakhir setelah menghitung *information gain,* sedangkan pada algoritma C5.0, perhitungan dilanjutkan dengan menghitung *gain ratio* (Apriyadi et al., 2022). Berikut adalah rumus untuk menghitung nilai *Entropy:*

$$E(S) = \sum_{i=1}^{n} - Pi * log_2 Pi \quad (1)$$

Keterangan :
S : Himpunan Kasus
n : Jumlah partisi S
Pi : Proporsi dari Si terhadap S
E:Entropy

Rumus untuk mencari nilai *Information Gain* :

$$IG(S,A) = E(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * E(Si) \quad (2)$$

Keterangan :
S : Himpunan kasus
A : Atribut
n : Jmlah partisi atribut A
| Si | : Jumlah kasus pada partisi ke i
| S | : Jumlah kasus dalam S
E:Entropy

Rumus untuk mencari nilai gain ratio :

$$Gain\ Ratio = \frac{Information\ Gain\ (S,A)}{\sum_{i=1}^{n} Entropy(Si)} \quad (3)$$

Keterangan:
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i

Langkah ini bertujuan untuk melakukan analisis dan pelatihan pada data latih menggunakan algoritma C5.0 untuk menghasilkan model yang dapat memprediksi atau mengklasifikasikan data uji dengan akurat.

2.5. Interpretation/Evaluation
Evaluasi merupakan tahap penting dalam pengembangan sebuah sistem, karena melalui evaluasi, pengguna dapat mengetahui tingkat akurasi, presisi, dan kesalahan *(error)* dari sistem yang dikembangkan. Salah satu teknik yang digunakan untuk mengevaluasi performa program klasifikasi obesitas adalah Confusion Matrix. Confusion Matrix membantu mengidentifikasi nilai-nilai penting seperti akurasi, presisi, recall, dan tingkat kesalahan dari sistem, yang dijelaskan melalui Tabel 1(Markoulidakis et al., 2021).

Tabel 1.*Confusion Matrix*

Kelas Prediksi			
		Kelas Positif	Kelas Negatif
Kelas Aktual	Kelas Positif	TP (True Positive)	FP (False Positive)
	Kelas Negatif	FN (False Negative)	TN (True Negative)

Sumber: Hasil penelitian (2024)

Confusion Matrix mencakup kelas prediksi dan kelas aktual, yang menunjukkan kinerja klasifikasi berdasarkan hasil prediksi dan data aktual. Matriks ini menampilkan jumlah data yang diklasifikasikan dengan benar dan salah untuk setiap kategori. Berdasarkan struktur *confusion matrix* ini, kita dapat dengan mudah menghitung nilai akurasi,

recall, dan presisi dari sistem yang dievaluasi. Berikut adalah rumus perhitungannya:

$$Clasification\ Akurasi = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100\% \quad (4)$$

Keterangan:
TP = jumlah data positif yang diklasifikasi benar
TN = jumlah data negatif yang diklasifikasi benar
FP = jumlah data positif yang diprediksi salah
FN = jumlah data negatif yang diprediksi salah

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp}+\text{Fp}} \times 100\% \quad (5)$$

Keterangan:
TP = jumlah data positif yang diklasifikasi benar
FP = jumlah data positif yang diprediksi salah

$$\text{Akurasi} = \frac{\text{Tp}}{\text{Tp}+\text{Fn}} \times 100\% \quad (6)$$

Keterangan:
FP = jumlah data positif yang diprediksi salah
FN = jumlah data negatif yang diprediksi salah

Confusion Matrix merupakan metode evaluasi yang sangat populer dalam menilai kinerja model klasifikasi, terutama dengan mengukur akurasi dan recall (Heydarian & Doyle, 2022). Analisis hasil dievaluasi menggunakan *confusion matrix*, yang terdiri dari notasi *Tp, Tn, Fp,* dan *Fn. Tp (True Positive)* merujuk pada jumlah kelas positif yang berhasil diklasifikasikan dengan benar sebagai positif. *Tn (True Negative)* mengindikasikan jumlah kelas negatif yang benar-benar diidentifikasi sebagai negatif. *Fp (False Positive)* menunjukkan jumlah kelas negatif yang salah diklasifikasikan sebagai positif, sementara *Fn (False Negative)* adalah jumlah kelas positif yang salah diidentifikasi sebagai negatif.

III. HASIL DAN PEMBAHASAN

Bagian hasil dan pembahasan menjelaskan temuan dari penelitian mengenai obesitas dan penerapan algoritma C5.0 pada dataset. Pada bagian ini, penjelasan dimulai dari proses pengumpulan data, dilanjutkan dengan tahap pra-pemrosesan data, hasil klasifikasi obesitas menggunakan algoritma C5.0, hingga evaluasi terhadap hasil yang dicapai dalam penelitian ini. Uraian berikut memberikan rincian mengenai hasil dan pembahasan dari setiap tahapan yang telah dilakukan dalam penelitian ini.

3.1. Data Selection

Penelitian ini menggunakan data publik yang diambil dari situs Kaggle (https://www.kaggle.com/datasets/ankurbajaj9/obesity), di mana file tersebut terdiri dari 2.111 data dengan 17 atribut yang tersedia dalam format *Comma Separated Values* (CSV). Salah satu atribut yang menjadi fokus utama sebagai label dalam penelitian ini adalah 'Class'. Dataset ini menjadi dasar dalam pengembangan dan evaluasi model untuk analisis klasifikasi obesitas yang dilakukan dalam penelitian ini.

Tabel 2.Deskripsi set data yang dikumpulkan

Atribut		Singkatan	Keterangan	
1.	Jenis belamin	Gender	Male, Female	
2.	Umur	Age	Integer Values	Numeric
3.	Tinggi badan	Height	Integer Values (Mt)	Numeric
4.	Berat badan	Weight	Integer Values (Kg)	Numeric
5.	Riwayat keluarga dengan Kelebihan berat badan	FHWO	Yes, No	
6.	Frekuensi konsumsi makanan berkalori tinggi	FAVC	Yes, No	
7.	Frekuensi konsumsi sayuran	FCVC	1: Always 2: Sometimes 3: Rarely	
8.	Jumlah makanan utama	NCP	1 to 2: UD 3: TR More than 3: MT	
9.	Konsumsi makanan di antara waktu makan	CAEC	S: Always CS: Usually A: Sometimes CN: Rarely	
10.	Merokok	SMOKE	Yes, No	
11.	Konsumsi air harian	CH2O	1: Less than one-liter 2: Between 1- and 2-liters 3: More than 2 liters	
12.	Atribut yang berhubungan dengan kondisi fisik (Pemantauan konsumsi kalori)	SCC	Yes, No	
13.	Frekuensi aktivitas fisik	FAF	1: 1 to 2 days 2: 3 to 4 days 3: 5 to 6 days 0: No physical activity	
14.	Waktu menggunakan perangkat teknologi	TUE	0: 0 to 2 hours 1: 3 to 5 hours 2: More than 5 hours	

15.	Konsumsi alcohol	CALC	NO: No consume of alcohol CF: Rarely S: Weekly D: Daily
16.	Transportasi yang digunakan	MTRANS	TP: Public transportation MTA: Motorbike BTA: Bike CA: Walking AU: Automobile
17.	Tingkat Obesitas	NObeyesdad	Insufficient_Weight Normal_Weight Overweight Overweight_Level_I Obesity_Type_I ---Obesity_Type_II Obesity_Type_III

Sumber: Hasil penelitian (2024)

Selanjutnya dari jumlah data 2.111, data dibagi menjadi dua dengan presentase 70% data training dan 30% data testing. 70% dari jumlah data 2.111 didapatkan data training dengan jumlah 1.477 Data training tersebutlah yang akan digunakan untuk mengimplementasikan algoritma C5.0 pada klasifikasi obesitas. dan 30% dari jumlah data 2.111 didapatkan data testing dengan jumlah 634 Data *testing* ini nantinya akan berfungsi untuk menguji performa dari algoritma C5.0 yang penulis gunakan, dengan metode pengujian *confusion matrix*.

3.2. Prepocesing Data

Pada tahap ini, penanganan terhadap nilai yang hilang (missing value) akan dilakukan, namun pada dataset yang digunakan tidak ditemukan adanya nilai yang hilang. Selanjutnya, akan dilakukan seleksi atribut dari 17 atribut yang terdapat dalam dataset obesitas yaitu *Gender, Age, Height, Weight,* FHWO, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS, NObeyesdad. namun atribut yang akan digunakan pada penelitian ini hanya 6 atribut yaitu *Weight*, FHWO, CAEC, *Height, Age* dan NObeyesdad Atribut yang diplih adalah atribut yang berpengaruh terhadap gejala-gejala yang paling dialami oleh penderita obesitas.

Tabel 3. Preprocessing Obesitas

Age	Height	Weight	FHWO	CAEC	NObeyesdad
21.0	1.62	64.0	Yes	Sometimes	Normal_Weight
23.0	1.52	56.0	Yes	Sometimes	Normal_Weight
27.0	1.80	77.0	Yes	Sometimes	Normal_Weight
22.0	1.80	87.0	No	Sometimes	Overweight_Level_I
22.0	1.78	89.8	No	Sometimes	Overweight_Level_II

Sumber: Hasil penelitian (2024)

Tabel 3 merupakan data Obesitas setelah dilakukan *preprocessing* pemilihan 6 atribut.

3.3 Transformation

Langkah selanjutnya adalah melakukan Transformasi Data, Transformasi Data digunakan untuk merubah data agar dapat dibaca oleh algoritma. Pada dataset Obesitas dilakukan transformasi pada atribut FHWO, CAEC, Nobeyesdad, yang memiliki data kategorikal menjadi numerik agar dapat di baca oleh algoritma.

Tabel 4. *Transformation* Obesitas

Age	Height	Weight	FHWO	CAEC	NObeyesdad
-0.522124	-0.875589	-0.862558	0.472291	0.300346	1
-0.522124	-1.947.599	-1.168.077	0.472291	0.300346	1
-0.206889	1.054.029	-0.366090	0.472291	0.300346	1
0.423582	1.054.029	0.015808	-2.117.337	0.300346	5
0.364507	0.839627	0.122740	-2.117.337	0.300346	6

Sumber: Hasil penelitian (2024)

Tabel 4 merupakan data Obesitas setelah dilakukan transformasi data.

3.4 Data Mining

Setelah melakukan beberapa tahapan yang diantaranya data *selection, preprocessing,* dan *transformation* maka tahap selanjutnya adalah data mining. Teknik pemodelan data mining yang dipilih adalah klasifikasi menggunakan algoritam C5.0. Pada proses pembentukan pohon klasifikasi algoritma C5.0 tahap pertama yaitu menentukan *node* akar, kemudian dilanjutkan dengan penentuan cabang untuk masing-masing *node*. Selanjutnya dilakukan pembagian kelas pada cabang yang telah diperoleh dan proses tersebut diulang hingga setiap cabang memiliki kelas. Adapun data yang digunakan untuk proses pembentukan pohon klasifikasi yaitu 70% dari keseluruhan data yakni sebanyak 1.477 sampel (*data training*), sedangkan sisanya 30% dari keseluruhan data yakni sebanyak 634 sampel digunakan sebagai data *testing* untuk pohon klasifikasi yang telah terbentuk.

Langkah awal dalam membentuk pohon keputusan adalah menghitung nilai *entropy* untuk setiap gejala (atribut) yang ada, sehingga perhitungan *information gain* dapat dilakukan berdasarkan hasil perhitungan *entropy* secara keseluruhan. Hasil perhitungan *entropy* dapat dilihat pada Tabel 5.

Tabel 5. *Transformation Entropy*

Atribut	Entropy
Age	87.062.795
Height	95.179.005
Weight	94.442.194
FHWO	0.6827028
CAEC	0.8373910

Sumber: Hasil penelitian (2024)

Langkah berikutnya adalah menghitung *information gain* untuk setiap atribut, yang kemudian akan digunakan dalam perhitungan *gain ratio* pada tahap selanjutnya. Hasil perhitungan *information gain* untuk semua atribut dapat dilihat pada Tabel 6 di bawah ini.

Tabel 6. Hasil Perhitungan *Information Gain*

Atribut	Entropy	Information Gain
Age	8.7062795	2.2332718
Height	9.5179005	2.4052114
Weight	9.4442194	2.5293750
FHWO	0.6827028	0.2328000
CAEC	0.8373910	0.2560776

Sumber: Hasil penelitian (2024)

Tahap terakhir adalah menghitung nilai dari gain ratio yang akan menjadi acuan pada penentuan *node* pada proses pembuatan *decision tree* dengan algoritma C5.0.

Tabel 7. Hasil Perhitungan *Gain Ratio*

Atribut	Entropy	Information Gain	Gain Ratio
Age	8.7062795	2.2332718	0.2565128
Height	9.5179005	2.4052114	0.2527040
Weight	9.4442194	2.5293750	0.2648226
FHWO	0.6827028	0.2328000	0.3409975
CAEC	0.8373910	0.2560776	0.2991242

Sumber: Hasil penelitian (2024)

Jika diperhatikan pada Tabel 7, nilai *Gain Ratio* tertinggi adalah pada atribut FHWO. Namun, atribut *Weight* memiliki *Information Gain* yang lebih tinggi, yang menunjukkan bahwa atribut ini memberikan pengurangan ketidakpastian terbesar dalam klasifikasi data. Oleh karena itu, berdasarkan algoritma C5.0, atribut Weight dipilih sebagai *node* akar (*root node*) karena memberikan informasi yang paling signifikan untuk memulai pembentukan pohon keputusan, meskipun FHWO memiliki *Gain Ratio* yang lebih tinggi.

Sumber: Hasil penelitian (2024)

Gambar 2 Decision Tree

Setelah melalui tahap perhitungan keseluruhan *node* atau biasa disebut dengan *gase training* dengan *decision tree*(Johnson & Khoshgoftaar, 2019). maka terbentuklah sebuah pohon keputusan yang dapat dilihat pada Gambar 2. Pohon keputusan ini merepresentasikan pola prediksi dan klasifikasi tingkat obesitas berdasarkan atribut-atribut yang telah dipilih. Setiap pola yang terbentuk menunjukkan hubungan antar gejala (atribut) yang mengacu pada keputusan akhir di *leaf node* untuk menentukan kategori tingkat obesitas seseorang.

Dari pohon keputusan yang terbentuk, beberapa kesimpulan dapat diambil:

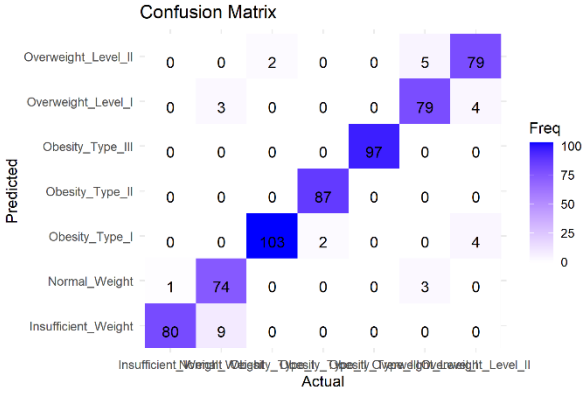
1. Berat Badan (*Weight* > 99.53 kg) Sebagai *Node* Akar:
   - Atribut *Weight* dipilih sebagai *node* akar karena memiliki *Information Gain* tertinggi, yang berarti atribut ini memberikan kontribusi terbesar dalam mengurangi ketidakpastian dalam dataset. Ketika berat badan seseorang lebih dari 99.53 kg, pohon keputusan lebih lanjut mengevaluasi atribut lain seperti usia dan tinggi badan untuk menentukan tingkat obesitas.
2. Cabang Berdasarkan Usia dan Tinggi Badan:
   - Jika seseorang memiliki berat badan lebih dari 99.53 kg dan usia lebih dari 26.04 tahun, tinggi badan menjadi faktor penentu. Orang dengan tinggi lebih dari 1.753 meter kemungkinan besar akan diklasifikasikan sebagai *Obesity Type II*. Namun, jika tingginya kurang dari

1.753 meter, kemungkinan besar mereka akan diklasifikasikan sebagai *Obesity Type I.*
- Di sisi lain, jika berat badan lebih dari 99.53 kg tetapi usia kurang dari 26.04 tahun, atribut lain seperti *Height* dan *Weight* akan menentukan apakah seseorang berada dalam kategori *Overweight Level I, Obesity Type I,* atau *Obesity Type II*

3. Cabang Berdasarkan Riwayat Keluarga dengan Kelebihan Berat Badan (FHWO):
   - Jika seseorang berada pada *Overweight Level I* tetapi memiliki riwayat keluarga dengan kelebihan berat badan, maka mereka kemungkinan akan diklasifikasikan lebih lanjut sebagai Overweight Level II
   - Sebaliknya, jika tidak ada riwayat keluarga dengan kelebihan berat badan, mereka cenderung tetap berada dalam kategori Overweight Level I.
4. Pola untuk Berat Badan <= 60 kg:
   - Jika berat badan seseorang kurang dari atau sama dengan 60 kg, pohon keputusan lebih lanjut mengevaluasi tinggi badan dan frekuensi konsumsi makanan (CAEC) untuk menentukan apakah seseorang memiliki berat badan normal atau kekurangan berat badan *(Insufficient Weight).*
   - Jika tinggi badan lebih dari 1.66 meter dan pola makan mereka hanya kadang-kadang, mereka mungkin diklasifikasikan sebagai memiliki berat badan normal. Namun, jika tinggi badan kurang dari 1.66 meter, mereka lebih mungkin memiliki Insufficient Weight.
5. Atribut Lain dalam Pohon Keputusan:
   - Atribut seperti CAEC (Frekuensi Konsumsi Makanan) dan FHWO (Riwayat Keluarga dengan Kelebihan Berat Badan) memainkan peran penting dalam cabang yang lebih rendah di pohon, yang menunjukkan bahwa meskipun kontribusi mereka tidak sebesar atribut utama seperti Weight dan Height, mereka tetap penting untuk klasifikasi akhir.

3.5. Interpretation/Evaluation

Penelitian ini menggunakan teknik confusion matrix untuk mengevaluasi kinerja model prediksi dan klasifikasi tingkat obesitas. Teknik ini memungkinkan pengukuran performa klasifikasi dengan cara membandingkan nilai prediksi dan nilai aktual(Quinlan, 1994). Dengan menggunakan bahasa pemrograman R, model prediksi dibuat menggunakan algoritma C5.0, dan evaluasi dilakukan dengan confusion matrix untuk menentukan akurasi, precision, recall, dan metrik lainnya.



Sumber: Hasil penelitian (2024)

Gambar 3 Confusion Matrix

Gambar 3 menunjukkan *Confusion Matrix* dari hasil klasifikasi tingkat obesitas menggunakan algoritma C5.0. Pada *Confusion Matrix* ini, baris merepresentasikan prediksi model, sedangkan kolom merepresentasikan nilai aktual dari data uji. Warna pada sel-sel matriks menunjukkan frekuensi prediksi, dengan warna yang lebih gelap menunjukkan jumlah prediksi yang lebih tinggi. Hasil ini menunjukkan bahwa model memiliki akurasi tinggi, dengan sebagian besar prediksi benar berada di diagonal utama matriks. Beberapa kesalahan terjadi, misalnya, beberapa prediksi *"Insufficient Weight*" salah diklasifikasikan sebagai "*Normal Weight",* tetapi secara keseluruhan, model menunjukkan performa yang sangat baik dalam mendeteksi tingkat obesitas.

```
Overall Statistics

               Accuracy : 0.9478
                 95% CI : (0.9274, 0.9638)
    No Information Rate : 0.1661
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.939

 Mcnemar's Test P-Value : NA
```

Sumber: Hasil penelitian (2024)

Gambar 4 Overall Statistics

Gambar 4 menampilkan statistik keseluruhan dari hasil evaluasi *Confusion Matrix*. Di sini, ditunjukkan bahwa akurasi keseluruhan model adalah 94.78%, yang berarti bahwa model berhasil memprediksi dengan benar hampir 95% dari seluruh data uji. Interval kepercayaan 95% (95% CI) berada pada rentang 92.74% hingga 96.38%, menunjukkan bahwa hasil ini dapat diandalkan. Nilai Kappa yang tinggi, yaitu 0.939, menunjukkan adanya kesepakatan yang sangat baik antara prediksi model dan nilai aktual, jauh lebih baik daripada jika prediksi dilakukan secara acak (ditunjukkan oleh *No Information Rate* yang rendah).

Tabel 8. Statistik Keseluruhan Evaluasi Model

Kelas	Sensitivitas	Spesifisitas	Nilai Prediktif Positif	Nilai Prediktif Negatif	Akurasi Seimbang	F1-Score
Insufficient Weight	98.77%	98.37%	89.89%	99.82%	98.57%	94.13%
Normal Weight	86.05%	99.27%	94.87%	97.83%	92.66%	90.26%
Obesity Type I	98.10%	98.86%	94.50%	99.62%	98.48%	96.26%
Obesity Type II	97.75%	100.00%	100.00%	99.63%	98.88%	98.86%
Obesity Type III	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Overweight Level I	90.80%	98.72%	91.86%	98.53%	94.76%	91.33%
Overweight Level II	90.80%	98.72%	91.86%	98.53%	94.76%	91.33%

Sumber: Hasil penelitian (2024)

Tabel 8 merangkum metrik evaluasi penting untuk setiap kelas obesitas, termasuk Sensitivitas *(Recall*), Spesifisitas, Nilai Prediktif Positif *(Precision),* Nilai Prediktif Negatif, Akurasi Seimbang, dan *F1-Score*. *F1-Score*, yang merupakan rata-rata harmoni antara *precision* dan *recall*, memberikan gambaran yang lebih komprehensif tentang keseimbangan performa model dalam mendeteksi positif sebenarnya sambil menghindari prediksi positif palsu. Dari tabel ini, dapat dilihat bahwa model memiliki performa yang sangat baik di hampir semua kelas, dengan nilai sensitivitas, spesifisitas, dan *F1-Score* yang tinggi. Kelas *"Obesity Type III"* memiliki nilai akurasi dan *F1-Score* yang sempurna, menunjukkan prediksi yang sangat akurat dan konsisten dalam kategori ini. Sementara itu, kelas *"Normal Weight"* sedikit lebih rendah dibandingkan kelas lainnya, yang menunjukkan bahwa model sedikit kesulitan dalam mengklasifikasikan individu dengan berat badan normal, namun tetap mempertahankan performa yang cukup baik.

IV. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, diperoleh kesimpulan bahwa algoritma C5.0 menunjukkan performa yang sangat baik dalam mengklasifikasikan tingkat obesitas menggunakan dataset dari Kaggle. Dengan proporsi data 70:30 untuk training dan testing, model ini menghasilkan akurasi sebesar 94,78%. Evaluasi menggunakan *confusion matrix* menunjukkan bahwa algoritma C5.0 memiliki *precision, recall*, dan *F1-Score* yang tinggi pada hampir semua kategori obesitas, terutama pada *Obesity Type III*, di mana model mencapai *precision* dan *recall* sempurna sebesar 100%. Atribut yang paling signifikan dalam klasifikasi ini adalah berat badan, diikuti oleh tinggi badan dan usia, yang menegaskan pentingnya faktor-faktor fisik dalam deteksi obesitas. Secara keseluruhan, algoritma C5.0 terbukti menjadi alat yang efektif dan andal dalam mendeteksi risiko obesitas, dan dapat digunakan sebagai dasar untuk pengembangan sistem pendukung keputusan di bidang kesehatan.

V. REFERENSI

Alexander Halim Santoso, Marcella E. Rumawas, David Limanan, & Freddy Ciptono. (2023). Penapisan Hiperuresemia dan Obesitas Pada Remaja di Jakarta Barat. *KREATIF: Jurnal Pengabdian Masyarakat Nusantara*, *3*(2), 121–128. https://doi.org/10.55606/kreatif.v3i2.1522

Alpiansah, A. B., & Ramdhani, Y. (2023). Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest Feature Optimization with Forward Selection on Obesity Rate Estimation using Random Forest. *SISTEMASI: Jurnal Sistem Informasi*, *12*(September), 860–873. http://sistemasi.ftik.unisi.ac.id

Amalda, R. N., Millah, N., & Fitria, I. (2022). Implementasi Algoritma C5.0 Dalam Menganalisa Kelayakan Penerima Keringanan Ukt Mahasiswa Itk. *Teorema: Teori Dan Riset Matematika*, *7*(1), 101. https://doi.org/10.25157/teorema.v7i1.6692

Apriyadi, A., Lubis, M. R., & Damanik, B. E. (2022). Penerapan Algoritma C5.0 Dalam Menentukan Tingkat Pemahaman Mahasiswa Terhadap Pembelajaran Daring. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *11*(1), 11–20. https://doi.org/10.34010/komputa.v11i1.7386

Benediktus, N., & Oetama, R. S. (2020). Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa Natanael. *Ultimatics : Jurnal Teknik Informatika*, *12*(1), 14–19.

Firmansyah, F., & Nurdiawan, O. (2023). Penerapan Data Mining Menggunakan Algoritma Frequent Pattern - Growth Untuk Menentukan Pola Pembelian Produk Chemicals. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 547–551. https://doi.org/10.36040/jati.v7i1.6371

Fitrianah, D., Gunawan, W., & Puspita Sari, A. (2022). Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction. *Februari*, *21*(1), 1–11.

Heydarian, M., & Doyle, T. E. (2022). *MLCM : Multi-Label Confusion Matrix*. 19083–19095.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0192-5

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*, *9*(4). https://doi.org/10.3390/technologies9040081

Quinlan, J. R. (1994). Book Review : C4 . 5 : Programs for Machine Learning. *Machine Learning*, *240*, 235–240.

Susindra, Y., & Permatasari, R. A. W. (2023). Pengaruh Media Pembelajaran Infografis Berbasis Aplikasi Android Terhadap Tingkat Pengetahuan Mengenai Obesitas Pada Remaja Putri. *ARTERI : Jurnal Ilmu Kesehatan*, *4*(2), 81–86. https://doi.org/10.37148/arteri.v4i2.269

Utomo, D. P., Sirait, P., & Yunis, R. (2020). Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0. *Jurnal Media Informatika Budidarma*, *4*(4), 994–1006. https://doi.org/10.30865/mib.v4i4.2355

Wijaya, A. C., Hasibuan, N. A., & Ramadhani, P. (2018). Implementasi Algoritma C5 . 0 Dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus : Kelurahan Mesjid Kecamatan Medan Kota). *Informasi Dan Teknologi Ilmiah (INTI)*, *13*, 192–198.

Zamasi, N. (2021). Implementasi Algoritma C 5 . 0 Pada Analisa Data Potensi Pertanian dan Perternakan. *TIN: Terapan Informatika Nusantara*, *2*(4), 184–190.