

IMPLEMENTASI DATA MINING DALAM MENGIDENTIFIKASI FAKTOR PASIEN YANG BERPOTENSI MENGALAMI OBESITAS MENGGUNAKAN ALGORITMA C4.5

Aghniya Nadira, Sri Farida Utami*

Bisnis Digital, Politeknik Sukabumi

Jl. Babakan Sirna No.25, Benteng, Kec. Warudoyong, Kota Sukabumi, Indonesia

srifaridautami@polteksmi.ac.id

ABSTRAK

Obesitas adalah penumpukan lemak secara berlebihan yang terjadi di daerah perut. Menurut *World Health Organization*, pada tahun 2022, 2,5 miliar orang dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan, angka ini akan terus meningkat jika masyarakat umum tidak menyadari gejala yang dapat menyebabkan penyakit obesitas dikemudian hari. Tujuan dari penelitian ini adalah untuk mengimplementasikan algoritma C4.5 dalam memprediksi penyakit obesitas, berdasarkan data yang telah diperoleh. Jumlah data yang digunakan adalah 1260 data, dimana 80% data merupakan data training dan 20% nya merupakan data testing, data tersebut memiliki 16 variabel yang relevan terkait obesitas. Berdasarkan hasil perhitungan dengan algoritma C4.5, atribut Weight menjadi akar utama pohon keputusan. Dari data yang telah di uji pada penelitian ini menghasilkan nilai akurasi, presisi, dan *recall* yang cukup besar, yaitu akurasi sebesar 98.80%, *precision* 98.48% dan *recall* 100%. Dari hasil penelitian, diketahui faktor-faktor utama yang dapat menyebabkan penyakit obesitas, yaitu faktor *weight* yang lebih dari 75,5kg, *family_history_with_overweight* dengan keterangan iya, dan SCC (Pemantauan Konsumsi Kalori) yang tidak dipantau. Dengan begitu, dapat dilakukan pencegahan agar kasus obesitas tidak cepat meningkat dengan melihat faktor-faktor penyebabnya.

Kata kunci : data mining, klasifikasi, pohon keputusan, algoritma C4.5

1. PENDAHULUAN

Obesitas menjadi penyebab utama berbagai penyakit yang terjadi baik di negara maju maupun berkembang, seperti penyakit jantung koroner (PJK), stroke iskemik, dan diabetes tipe 2 [1].

World Health Organization melaporkan bahwa pada tahun 2022, 2,5 miliar orang dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan, termasuk lebih dari 890 juta orang dewasa yang menderita obesitas. Hal ini setara dengan 43% orang dewasa berusia 18 tahun ke atas (43% pria dan 44% wanita) yang kelebihan berat badan; meningkat dari tahun 1990, ketika 25% orang dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan [2].

Obesitas adalah penumpukan lemak secara berlebihan yang terjadi di daerah perut. Meningkatnya populasi obesitas dikaitkan dengan perubahan pola makan dan gaya hidup tidak sehat. Adanya Perubahan perilaku pola makan yang tidak sehat, seperti pola makan yang mengkonsumsi lemak berlebih, bahan pengawet, dan rendah serat dapat memicu obesitas [3].

Untuk mengatasi masalah ini, obesitas dapat dicegah dengan menjalankan pola hidup sehat salah satunya dengan membatasi kandungan yang berlebihan pada komposisi makanan sehari-hari. Selain itu, obesitas dapat di minimalisir dengan olahraga ringan yang dapat dilakukan setiap hari. Untuk mengetahui seseorang mengidap obesitas atau tidak dapat dilakukan dengan menimbang berat badan dan mencatatnya secara berkala, dapat dilakukan sendiri ataupun konsultasi dengan dokter gizi.

Pencatatannya juga dapat menggunakan teknik klasifikasi dengan data mining.

Penelitian ini bertujuan untuk mengimplementasikan algoritma *Decision Tree* dalam data mining untuk mengklasifikasikan data obesitas. Implementasi ini menggunakan dataset yang terdiri dari atribut terkait obesitas yang relevan. Dengan memanfaatkan kemampuan algoritma C4.5 untuk memodelkan data dan mengambil keputusan, diharapkan dapat mengembangkan model yang dapat memprediksi obesitas pada seseorang secara akurat.

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data mining merupakan serangkaian proses untuk mengeksplorasi nilai tambah dari kumpulan data berupa pengetahuan yang tidak diketahui secara manual. Oleh karena itu, penambangan data menjadi semakin penting sebagai alat untuk mengubah data menjadi informasi [4].

Dalam beberapa tahun terakhir, data mining telah menarik perhatian industri informasi secara signifikan. Hal ini disebabkan oleh jumlah data yang sangat besar yang tersedia dan meningkatnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang bermanfaat [5].

Dalam penelitian ini, data mining yang digunakan merupakan metode klasifikasi. Klasifikasi adalah proses penambangan data yang bertujuan untuk menemukan pola berharga dalam jumlah data yang relatif besar hingga sangat besar. [6]

2.2. Decision Tree

Decision tree atau dapat disebut juga pohon keputusan merupakan metode klasifikasi yang cukup terkenal. Metode ini mengubah fakta penting menjadi pohon keputusan yang mewakili aturan dan membuat aturan tersebut mudah dipahami manusia [7].

Pohon keputusan juga digunakan untuk menggali data dan menemukan hubungan antara sekumpulan variabel input dan tujuan [8].

Decision tree terdiri dari sejumlah node yang saling terhubung melalui cabang. Ada tiga jenis node dalam *decision tree*:

1. Root node, merupakan node yang tidak memiliki input tetapi memiliki lebih dari satu output.
2. Internal node, memiliki satu input dan lebih dari dua output.
3. Leaf atau terminal node, memiliki satu input dan tidak memiliki output.

Pada setiap *decision tree*, setiap leaf memiliki nama kelas. Root node dan internal node mengandung aturan yang digunakan untuk membedakan data berdasarkan karakteristik yang berbeda [9].

2.3. Algoritma C4.5

Algoritma C4.5 merupakan algoritma untuk membangun *decision tree* atau pohon keputusan. Pohon keputusan adalah metode klasifikasi dan prediksi yang cukup terkenal. Pohon keputusan digunakan untuk menemukan hubungan antara sekumpulan variabel input dan variabel target [10]. Pohon keputusan dapat membagi kumpulan data besar menjadi kumpulan data yang lebih kecil dengan menerapkan serangkaian aturan keputusan [8].

2.4. Confusion Matrix

Confusion matrix adalah metode yang digunakan untuk menghitung tingkat akurasi dalam data mining [10].

Metode ini yang sering digunakan untuk mengukur nilai akurasi. Pengukuran kinerja menggunakan *confusion matrix* melibatkan empat istilah yang menggambarkan hasil klasifikasi. Keempat istilah tersebut adalah:

1. *False Positive* (FP), yaitu data negatif yang diprediksi sebagai data positif.
2. *False Negative* (FN), yaitu data positif yang diprediksi sebagai data negatif.
3. *True Positive* (TP), yaitu data positif yang diprediksi dengan benar.
4. *True Negative* (TN), yaitu data negatif yang diprediksi dengan benar. [11]

2.5. RapidMiner

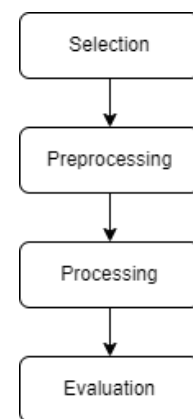
RapidMiner dikembangkan oleh perusahaan yang bernama sama dengan menyediakan lingkungan terintegrasi persiapan data, penambangan teks, dan analisis prediktif serta merupakan *software* ilmu data [12].

RapidMiner adalah alat penambangan data yang dikembangkan menggunakan bahasa pemrograman Java. Java dikenal memiliki banyak kelebihan, seperti kesederhanaan, keamanan, kekuatan, pengaruh, kemampuan berorientasi objek yang tinggi, serta berbagai keuntungan lainnya [13].

3. METODE PENELITIAN

Metode penelitian adalah serangkaian langkah dan waktu yang dapat dijadikan acuan dalam analisis data, serta merupakan rancangan kegiatan penelitian untuk menyelidiki, merumuskan, dan menganalisis data [14].

Secara garis besar penelitian ini akan melakukan perhitungan terhadap dataset dengan menggunakan algoritma klasifikasi C4.5. Adapun tahapan-tahapan yang digunakan pada penelitian ini, sebagai berikut:



Gambar 1. Tahapan metode penelitian

3.1. Selection

Data yang digunakan dalam penelitian ini berasal dari Kaggle, yaitu Obesitas dataset oleh Cahya Alkahfi. Dataset yang digunakan terdiri dari 1260 *record* data dengan 16 variabel termasuk satu variabel *Outcome* yang terdiri dari dua kelas keputusan: Normal dan *Obesity*.

3.2. Preprocessing

Tahap ini penting karena data yang terstruktur dengan baik akan berdampak positif pada hasil analisis dan pengambilan keputusan. Selama tahap prapemrosesan data penelitian ini, beberapa prosedur dilakukan, seperti pembersihan data, pengurangan data, dan normalisasi data [15].

3.3. Processing

Karena jumlah data yang digunakan cukup banyak, proses analisis data mining dilakukan secara manual untuk mempermudah pemahaman tentang penggunaan teknik klasifikasi dengan algoritma C4.5. Rumus Algoritma C4.5 terbagi menjadi dua rumus, yang pertama yaitu rumus untuk mencari nilai entropy. Entropy digunakan untuk memilih atribut yang paling optimal agar tidak menghasilkan cabang lebih lanjut. Adapun rumus entropy adalah sebagai berikut:

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

Keterangan:

S = himpunan kasus

n = jumlah partisi S

pi = probabilitas yang didapat dari kasus I dibagi total kasus.

Dan kedua, gain digunakan untuk mengukur apakah atribut yang dipilih sesuai untuk memecahkan node [16]. Untuk rumus kedua sebagai berikut :

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

Keterangan :

S = himpunan kasus

A = atribut

n = jumlah partisi atribut

|S_i| = jumlah kasus pada partisi ke-i

|S| = jumlah kasus dalam S

3.4. Evaluation

Setelah hasil klasifikasi diperoleh, kemudian hasilnya dievaluasi menggunakan *cross-validation* (*confusion matrix*) untuk mengetahui akurasi, presisi, dan recall dari hasil model yang digunakan [4]. Pengujian dilakukan untuk menentukan tingkat akurasi dari pohon keputusan algoritma C4.5. dalam pengujian ini dataset dibagi menjadi dua bagian, yaitu sebanyak 80 % untuk dijadikan sebagai data *training* dan sebanyak 20 % dijadikan sebagai data *testing*.

4. HASIL DAN PEMBAHASAN

4.1. Selection

Untuk melakukan penelitian ini, data diperoleh dari website Kaggle. Dataset yang digunakan di upload oleh akun yang bernama Cahya Alkahfi dengan file CSV bernama obesitas.

4.2. Preprocessing

Pada penelitian ini menggunakan *Software* RapidMiner 10.3 untuk menganalisis data dengan algoritma C4.5 sebagai pemodelan. Dari banyaknya data yang tersedia, hanya tersisa 1260 data yang digunakan untuk penelitian ini. Pengolahan data obesitas terdiri dari 16 variabel yang disajikan pada tabel 1.

Tabel 1. Variabel data penelitian

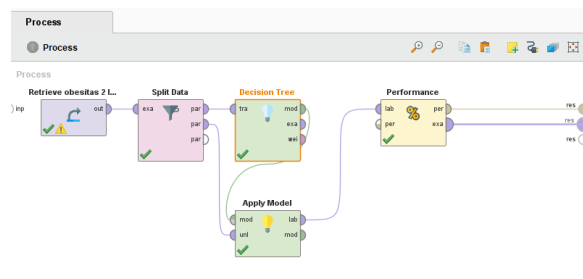
No	Variabel
1	Gender
2	Age
3	Height
4	Weight
5	Family_history_With_Overweight
6	FAVC
7	FCVC
8	NCP
9	CAEC
10	Smoke
11	CH2O

No	Variabel
12	SCC
13	FAF
14	CALC
15	MTrans
16	Obeyesdad

Selanjutnya dari jumlah data 1260, data dibagi menjadi dua dengan persentase 80% data *training* dan 20% data *testing*. 80% dari jumlah data 1260 didapatkan data *training* dengan jumlah 1009 data dan 20% dari jumlah data 1260 didapatkan data *testing* dengan jumlah 251 data.

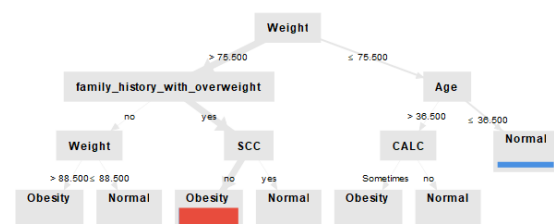
4.3. Processing

Dari data yang peroleh selanjutnya dibentuk pohon keputusan menggunakan algoritma C4.5 untuk menghasilkan alur keputusan. Berikut ini merupakan proses klasifikasi data obesitas menggunakan RapidMiner 10.3 :



Gambar 2. Proses klasifikasi menggunakan *software* RapidMiner

Setelah dilakukan proses pada Gambar 2 dengan metode algoritma C4.5, maka terbentuk pohon keputusan seperti gambar di bawah ini:



Gambar 3. *Decision tree* dari hasil algoritma C4.5

Berdasarkan *Decision tree* untuk model klasifikasi obesitas diatas. Dapat dilihat bahwa atribut yang mempunyai pengaruh utama dalam penentuan obesitas adalah atribut *weight*, kemudian *family_history_with_overweight*, *age*, SCC dan terakhir CALC. Dengan begitu menghasilkan *rule-rule* sebagai berikut :

Tree

```

Weight > 75.500
| family_history_with_overweight = no
| | Weight > 88.500: Obesity {Normal=0, Obesity=5}
| | Weight ≤ 88.500: Normal {Normal=7, Obesity=2}
| family_history_with_overweight = yes
| | SCC = no: Obesity {Normal=6, Obesity=767}
| | SCC = yes: Normal {Normal=3, Obesity=2}
Weight ≤ 75.500
| Age > 36.500
| | CALC = Sometimes: Obesity {Normal=0, Obesity=2}
| | CALC = no: Normal {Normal=4, Obesity=0}
| Age ≤ 36.500: Normal {Normal=210, Obesity=0}

```

Gambar 4. Rule dari hasil decision tree

Rule-rule yang dihasilkan dari decision tree hasil pengujian dideskripsikan sebagai berikut :

- Jika *Weight* > 75,5 dan *family_history_with_overweight* = No AND *Weight* > 88.5 maka *Obesity*
- Jika *Weight* > 75,5 dan *family_history_with_overweight* = No dan *Weight* ≤ 88.5 maka *Normal*
- Jika *Weight* > 75,5 dan *family_history_with_overweight* = Yes dan *SCC* = No maka *Obesity*
- Jika *Weight* > 75,5 dan *family_history_with_overweight* = Yes dan *SCC* = Yes maka *Normal*
- Jika *Weight* < 75,5 dan *age* > 36 dan *CALC* = Sometime maka *Obesity*
- Jika *Weight* < 75,5 dan *age* > 36 dan *CALC* = No maka *Normal*
- Jika *Weight* < 75,5 dan *age* < 36 dan *CALC* = Sometime maka *Normal*

Keterangan :

Weight = berat badan

Famliy history with Overweight = riwayat keluarga

SCC = pemantauan konsumsi kalori

CALC = konsumsi alkohol

4.4. Evaluasi Model

Setelah pengujian untuk rule pohon keputusan dilanjutkan dengan pengujian data testing menggunakan *Software Rapid Miner*. Data testing yang diuji sebanyak 251 data. Didapatkan *Confusion Matrix* seperti dibawah ini :

accuracy: 98.80%

	true Normal	true Obesity	class precision
pred Normal	54	0	100.00%
pred Obesity	3	194	98.48%
class recall	94.74%	100.00%	

Gambar 5. Hasil confusion matrix dari hasil penelitian

Dari tabel di atas, memperoleh jumlah True Positive (TP) sebanyak 54 yang dianggap true dan sesuai dengan prediksinya, lalu False Positive (FP) terdapat sebanyak 3 dengan prediksi false ternyata

hasil klasifikasi true, True Negative (TN) sebanyak 194 diprediksi false sesuai dengan klasifikasinya. Dan untuk False Negative (FN) sebanyak 0 yang menurut prediksi true tapi hasilnya false. Persentase akurasi yang didapatkan menggunakan metode algoritma ini dapat dilihat dari gambar 4 sebesar 98.80%. dengan *presicision* sebesar 98.48% dan *recall* sebesar 100.00%. Dari hasil tersebut, menunjukkan bahwa algoritma C4.5 memberikan akurasi yang tinggi dalam mengklasifikasikan obesitas.

5. KESIMPULAN DAN SARAN

Dari hasil penelitian yang telah dilakukan bisa ditarik kesimpulan bahwa klasifikasi dengan metode algoritma C4.5 terhadap data klasifikasi obesitas menghasilkan *rule-rule* dan *knowledge* yang dapat digunakan untuk memprediksi tingkat obesitas pada tubuh seseorang. Dari data total yang berjumlah 1260 terbagi ke dalam dua data yang berbeda, yaitu data *training* sebanyak 1009 data dan data *testing* sebanyak 251 data, menghasilkan tingkat akurasi yang cukup tinggi dengan persentase sebesar 98.80%. selain tingkat akurasi didapatkan juga hasil *presicision* sebesar 98.48% dan hasil *recall* sebesar 100.00%.

Dari hasil penelitian diatas, dapat diketahui faktor-faktor utama yang bisa menyebabkan penyakit obesitas, yaitu faktor *weight* yang lebih dari 75,5kg, *family_history_with_overweight* dengan keterangan iya, dan *SCC* (Pemantauan Konsumsi Kalori) yang tidak dipantau. Dengan begitu, dapat dilakukan pencegahan agar kasus obesitas tidak cepat meningkat dengan melihat faktor-faktor penyebabnya.

DAFTAR PUSTAKA

- [1] B. R. K. Salim, D. M. Wihandani, and N. N. A. Dewi, "Obesitas sebagai faktor risiko terjadinya peningkatan kadar trigliserida dalam darah: tinjauan pustaka," *Intisari Sains Medis*, 2021, [Online]. Available: <https://mail1.isainsmedis.id/index.php/ism/article/view/1031>
- [2] WHO, "Obesity and overweight," *World Health Organizatio*, 2024. <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> (accessed Jun. 11, 2024).
- [3] D. Dewanti, A. Syauqy, E. R. Noer, and A. Pramono, "Hubungan Pola Makan dan Aktivitas Fisik Dengan Obesitas Sentral Pada Usia Lanjut di Indonesia: Data Riset Kesehatan Dasar," *Gizi Indones.*, 2022, [Online]. Available: https://www.persagi.org/ejournal/index.php/Gizi_Indon/article/view/662
- [4] A. H. Nasrullah, "Implementasi algoritma Decision Tree untuk klasifikasi produk laris," *J. Ilm. Ilmu Komput. Fak. ...*, 2021, [Online]. Available: <http://ejournal.fikom-unasman.ac.id/index.php/jikom/article/view/203>
- [5] S. F. Utami, "Penerapan Data Mining Algoritma Decision Tree Berbasis PSO," *Semin. Nas. Teknol. Komput. &Sains ...*, 2020, [Online].

- Available: <http://seminar-id.com/prosiding/index.php/sainteks/article/view/523>
- [6] H. Hafizan and A. N. Putri, "Penerapan Metode Klasifikasi Decision Tree Pada Status Gizi Balita Di Kabupaten Simalungun," *Kesatria J. Penerapan Sist.* ..., 2020, [Online]. Available: <http://www.pkm.tunasbangsa.ac.id/index.php/kesatria/article/view/23>
- [7] Y. Irawan, "Penerapan Algoritma Decision Tree C4. 5 Untuk Memprediksi Kelayakan Calon Pendorong Melakukan Donor Darah Dengan Klasifikasi Data Mining," *JTIM J. Teknol. Inf. dan Multimed.*, 2021, [Online]. Available: <https://journal.sekawan-org.id/index.php/jtim/article/view/75>
- [8] K. A. Saputra, J. T. Hardinata, M. R. Lubis, and ..., "Klasifikasi Algoritma C4. 5 dalam penerapan tingkat kepuasan Siswa terhadap media pembelajaran online," *KLIK Kaji. Ilm.* ..., 2020, [Online]. Available: <http://www.djournals.com/klik/article/view/120>
- [9] D. E. Sinaga, A. P. Windarto, and R. A. Nasution, "Analisis Data Mining Algoritma Decision Tree Pada Prediksi Persediaan Obat (Studi Kasus: Apotek Franch Farma)," *Klik Kaji. Ilm.* ..., 2022, [Online]. Available: <http://www.djournals.com/klik/article/view/328>
- [10] N. Nuradha, H. Hazriani, and Y. Yuyun, "Penerapan Algoritma C4. 5 dalam Mengidentifikasi Karakteristik Pasien Beresiko Diabetes," *Pros. SISFOTEK*, 2023, [Online]. Available: <http://seminar.iaii.or.id/index.php/SISFOTEK/article/view/412>
- [11] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4. 5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan ...)*, 2020, [Online]. Available: <https://jurnal.tau.ac.id/index.php/siskom-kb/article/view/173>
- [12] V. M. Paul, I. Gunawan, B. E. Damanik, and ..., "Penerapan Data Mining Menggunakan Algoritma C4. 5 Dalam Menentukan Kelayakan Penerimaan Bantuan Bedah Rumah Pada Desa Tiga Dolok," *J. Sos.* ..., 2021, [Online]. Available: <https://sostech.greenvest.co.id/index.php/sostech/article/view/102>
- [13] A. Al Arif, M. Firdaus, and Y. Maruhawa, "Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4. 5, Naïve Bayes, dan KNN: Comparison of Data Mining Methods for Prediction of ...," *SENTIMAS Semin. Nas.* ..., 2022, [Online]. Available: <https://journal.irpi.or.id/index.php/sentimas/article/view/308>
- [14] Y. Fitriani, S. Defit, and G. W. Nurcahyo, "Prediksi Hasil Belajar Siswa Secara Daring pada Masa Pandemi COVID-19 Menggunakan Metode C4. 5," *J. Sistim Inf. Dan ...*, 2021, [Online]. Available: <http://jsisfotek.org/index.php/JSisfotek/article/view/54>
- [15] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4. 5," *J. Comput. Inf. Syst.* ..., 2023, [Online]. Available: <https://www.journal-computing.org/index.php/journal-cisa/article/view/393>
- [16] D. Pascalina, R. Widhiastono, and C. Julianne, "Pengukuran Kesiapan Transformasi Digital Smart City Menggunakan Aplikasi Rapid Miner," *Technomedia J.*, vol. 7, no. 3, pp. 293–302, 2022, doi: 10.33050/tmj.v7i3.1914.