JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim/

Prediksi Persentase *Body Fat* Menggunakan Algoritma CART dan M5′

Uswatun Hasanah[1]* dan Ade Nurhopipah[2]

[1] Program Studi Teknik Komputer, Fakultas Teknik, Universitas Negeri Semarang; uswatun_hasanah@mail.unnes.ac.id
[2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto; ade_nurhopipah@amikompurwokerto.ac.id
* Korespondensi: uswatun_hasanah@mail.unnes.ac.id

Sitasi: Hasanah, U.; Nurhopipah, A. (2023). Prediksi Persentase Body Fat Menggunakan Algoritma CART dan M5′. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 4(4), 351-363. https://doi.org/10.35746/jtim.v4i4.316

Abstract: Body Fat Percentage (BFP) is a measurement of total body fat that is used as an accurate measurement for the diagnosis of obesity. BFP measurement is sometimes difficult and inconvenient to perform, even though the picture of BFP's value is very important for someone to find out the chances of being obese. To overcome this, data mining techniques can be used to measure the predictions of BFP values in a more practical way. This study implements data mining techniques, namely the CART and M5′ algorithm to predict a person's BFP value based on his/her body measurement. The CART algorithm uses the sample average values at leaf nodes to make numerical predictions, while the M5' algorithm builds a regression model for each leaf node with a hybrid approach. Regression trees provide a simple way of explaining the relationship between features and numerical results, but more complex model trees also provide more accurate results. In this study, the results show that the M5' algorithm is superior to the BFP dataset with a correlation value of 0.86 and an MAE value of 3.86.

Keywords: Body Fat Percentage; CART; M5′, Prediction

Abstrak: *Body Fat Percentage* (BFP) merupakan ukuran lemak tubuh total yang digunakan sebagai pengukuran yang akurat untuk diagnosis obesitas. Pengukuran BFP terkadang sulit dan tidak nyaman untuk dilakukan, padahal gambaran nilai BFP sangat penting bagi seseorang untuk mengetahui peluang dirinya terkena obesitas. Untuk mengatasi hal tersebut, teknik *data mining* dapat digunakan untuk mengukur prediksi nilai BFP dengan cara yang lebih praktis. Studi ini menerapkan teknik *data mining*, yaitu algoritma CART dan M5′ untuk memprediksi nilai BFP seseorang berdasarkan pengukuran tubuhnya. Algoritma CART menggunakan nilai rata-rata sampel pada simpul daun untuk membuat prediksi numerik, sedangkan algoritma M5′ membangun model regresi pada setiap simpul daun dengan pendekatan hibrid. Pohon regresi memberikan cara sederhana untuk menjelaskan hubungan antara fitur dan hasil numerik, tetapi pohon model yang lebih kompleks juga memberikan hasil yang lebih akurat. Pada penelitian ini, hasil menunjukkan bahwa algoritma M5′ lebih unggul pada dataset BFP dengan nilai korelasi sebesar 0,86 dan nilai MAE sebesar 3,86.

Kata kunci: Body Fat Percentage; CART; M5′, Prediksi.

1. Pendahuluan

World Health Organization (WHO) melaporkan bahwa kelebihan berat badan dan obesitas lebih sering terkait dengan kematian di seluruh dunia daripada kekurangan berat badan [1]. Pada tahun 2016, lebih dari 1,9 miliar orang dewasa berusia 18 tahun ke atas mengalami kelebihan berat badan. Dari jumlah tersebut lebih dari 650 juta orang dewasa mengalami obesitas. Pada tahun 2016, 39% orang dewasa berusia 18 tahun ke atas (39% pria dan 40% wanita) mengalami kelebihan berat badan. Secara keseluruhan, sekitar 13% populasi orang dewasa dunia (11% pria dan 15% wanita) mengalami obesitas pada tahun 2016. Prevalensi obesitas di seluruh dunia hampir tiga kali lipat antara tahun 1975 dan 2016 [1]. Penyebab mendasar dari obesitas dan kelebihan berat badan adalah ketidakseimbangan energi antara kalori yang dikonsumsi dan kalori yang dikeluarkan. Resiko kesehatan umum pada kasus kelebihan berat badan dan obesitas adalah penyakit kardiovaskular (terutama penyakit jantung dan stroke), diabetes, dan gangguan muskuloskeletal. Obesitas juga berkaitan erat dengan kelebihan lemak tubuh di mana kelebihan lemak tubuh dapat meningkatkan risiko enam jenis kanker, termasuk kanker usus, kerongkongan, pankreas, ginjal, endometrium, dan payudara [2]. Selain itu, diabetes tipe II juga ditemukan pada orang-orang yang memiliki terlalu banyak lemak tubuh. Oleh karena itu, menghindari obesitas dan mencegah kelebihan lemak tubuh menjadi isu yang sangat penting.

Body Fat Percentage (BFP) atau persentase lemak tubuh merupakan ukuran lemak tubuh total yang digunakan sebagai pengukuran yang akurat untuk diagnosis obesitas [3]. Beberapa teknik yang digunakan untuk mengestimasi nilai BFP adalah *anthropometry* (menggunakan massa tubuh, lingkar dan diameter, ketebalan lipatan kulit, dan lain-lain), *underwater weighing* (UWW), *dual energy X-ray absorptiometry* (DEXA), *bioelectrical impedance analysis* (BIA), *computed tomography* (CT), *magnetic resonance imaging* (MRI) dan *near infrared interactance* [4]–[6]. Meskipun demikian, teknik-teknik tersebut tidak selalu akurat dan nyaman digunakan [7], [8]. Sebagai gantinya, teknik *data mining* [9] dan *machine learning* [10] dapat digunakan untuk melakukan tugas prediksi atau klasifikasi dari suatu dataset. Tugas-tugas tersebut meliputi banyak hal, dari bidang akademik [11] sampai dengan bidang kesehatan. Pada bidang kesehatan, teknik *data mining* dan *machine learning* dapat digunakan untuk memprediksi persentase body fat pada tubuh seseorang. Prediksi persentase *Body Fat* dalam tubuh manusia penting untuk memberikan diagnosis awal sebelum akhirnya seseorang dapat mendatangi para ahli dan melakukan pemeriksaan menyeluruh terkait dengan prediksi yang mungkin terjadi. Apabila seseorang mengalami keterlambatan diagnosis persentase *Body Fat*, maka peluang untuk munculnya beberapa penyakit penyerta juga menjadi semakin besar sementara keberadaan penyakit-penyakit tersebut tidak disadari sejak awal.

Salah satu teknik data mining yang bisa digunakan untuk memprediksi BFP adalah metode regresi. Penelitian ini bertujuan untuk memprediksi BFP menggunakan metode pohon regresi dan pohon model dengan memanfaatkan *dataset* yang diusulkan oleh Johnson [8]. *Data set* yang digunakan terdiri dari nilai BFP yang ditentukan berdasarkan pengukuran penimbangan bawah air dan 13 pengukuran lingkar tubuh pada 252 laki-laki. Algoritma *Classification and Regression Tree* (CART) digunakan untuk memprediksi nilai BFP. Metode pohon regresi dipilih karena memiliki beberapa keunggulan, yaitu: memiliki pendekatan yang paling umum untuk memodelkan data numerik; dapat disesuaikan untuk memodelkan hampir semua data; serta memberikan perkiraan hubungan antara fitur dan hasilnya. Pada penelitian ini, algoritma CART dievaluasi dan algoritma M5 digunakan untuk meningkatkan performa prediksi.

2. Bahan dan Metode

2.1. Alat dan Bahan Penelitian

2.1.1. Alat

Alat yang digunakan pada penelitian ini adalah sebagai berikut:

- Laptop MacBook Air M1 (2020) 256 GB;
- R *programming* menggunakan RStudio 2022.07.2

2.1.2. Bahan

Bahan yang digunakan pada penelitian ini adalah *Body Fat Prediction Dataset* yang diusulkan oleh Johnson [8]. Dataset ini terdiri dari 252 pengukuran BFP berdasarkan pengukuran penimbangan bawah air dan 13 pengukuran lingkar tubuh. Dataset ini memiliki 15 variabel seperti yang ditunjukkan pada Tabel I. Variabel target (BFP) dinotasikan dengan Y dan pengukurannya melibatkan variabel Density (D) yang merupakan nilai kepadatan yang ditentukan melalui penimbangan bawah air. Persamaan Siri digunakan untuk mengukur nilai Y seperti yang ditunjukkan oleh persamaan (1).

$$Y = \frac{45}{D} - 450 \tag{1}$$

Tabel 1. Definisi variabel dalam dataset BFP

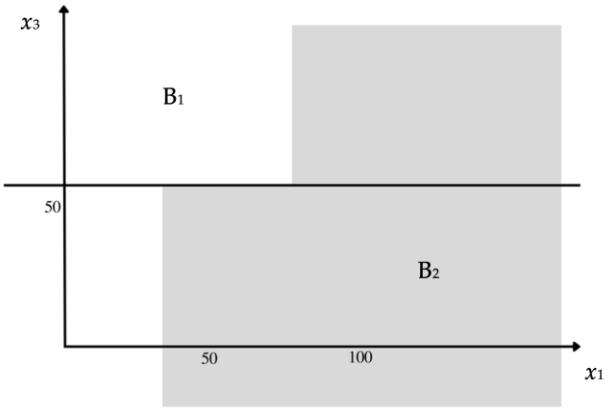
Variabel	Keterangan
D	Kepadatan yang ditentukan dari penimbangan bawah air
Y	BFP (*Body Fat Percentage*)
X_1	Umur (tahun)
X_2	Tinggi badan (cm)
X_3	Berat badan (kg)
X_4	Lingkar leher (cm)
X_5	Lingkar dada (cm)
X_6	Lingkar perut (cm)
X_7	Lingkar pinggul (cm)
X_8	Lingkar paha (cm)
X_9	Lingkar lutut (cm)
X_{10}	Lingkar pergelangan kaki (cm)
X_{11}	Lingkar bisep (memanjang) (cm)
X_{12}	Lingkar lengan bawah (cm)
X_{13}	Lingkar pergelangan tangan (cm)

2.2. Metode Penelitian

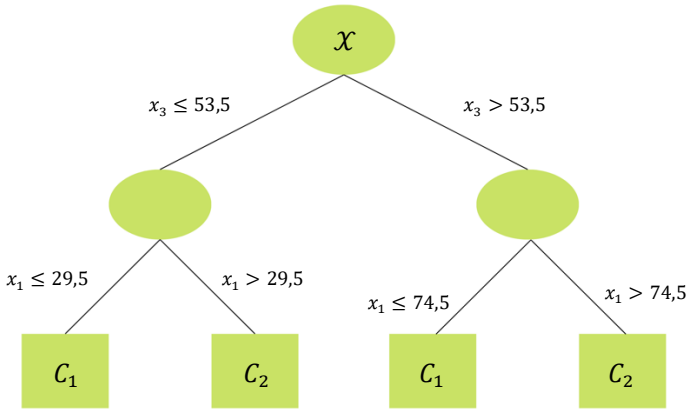
2.2.1. Algoritma CART

Pada tahun 1980-an, [12] mengembangkan algoritma CART (*Classification And Regression Trees*) yang digunakan untuk mencocokkan pohon pada data. Secara umum permasalahan klasifikasi adalah menempatkan setiap kasus pada sampel ke dalam satu dari sejumlah kelas yang memungkinkan. Suatu set pengukuran diberikan untuk suatu kasus, kemudian klasifikasi digunakan untuk memprediksi di kelas mana kasus tersebut berada. Pengklasifikasi adalah aturan yang menetapkan keanggotaan kelas yang diprediksi berdasarkan pengukuran terkait, $x_1, x_2, \dots, x_{K-1}$, dan x_K. Misalkan ruang pengukuran $\mathcal{X}$ adalah himpunan dari semua nilai $(x_1, \dots, x_K)$, dan $\mathcal{C} = \{c_1, c_2, \dots, c_J\}$ sebagai kelas-kelas yang memungkinkan, pengklasifikasi merupakan suatu fungsi dengan domain $\mathcal{X}$ dan kodomain $\mathcal{C}$, dan menghubungkan suatu partisi $\mathcal{X}$ dengan himpunan terpisah, $B_1, B_2, \dots, B_J$, sehingga kelas yang diprediksi adalah j jika $\mathbf{x} \in B_j$, di mana $\mathbf{x} = (x_1, \dots, x_K)$.

Pengklasifikasi pohon terstruktur dibangun dengan membuat pemisahan berulang dari subset $\mathcal{X}$, sehingga struktur hirarki bisa terbentuk. Sebagai contoh, $\mathcal{X}$ pertama kali dibagi menjadi $\{\mathbf{x} \mid x_3 \leq 53.5\}$ dan $\{x \mid x_3 > 53.5\}$. Selanjutnya himpunan pertama dapat dibagi lagi menjadi $A_1 = \{\mathbf{x} \mid x_3 \leq 53.5,\ x_1 \leq 29.5\}$ dan $A_2 = \{\mathbf{x} \mid x_3 \leq 53.5,\ x_1 > 29.5\}$, dan himpunan yang lainnya dapat dibagi menjadi $A_3 = \{\mathbf{x} \mid x_3 > 53.5,\ x_1 \leq 74.5\}$ dan $A_4 = \{\mathbf{x} \mid x_3 > 53.5,\ x_1 > 74.5\}$. Jika hanya terdapat dua kelas ($J = 2$), untuk kasus yang kelasnya tidak diketahui di mana $\mathbf{x}$ milik A_1 atau A_3 maka harus diklasifikasikan sebagai c_1 (diprediksi menjadi kelas c_1), dan kasus di mana $\mathbf{x}$ milik A_2 atau A_4 harus diklasifikasikan sebagai c_2. Sesuai dengan notasi di atas, maka $B_1 = A_1 \cup A_3$ dan $B_2 = A_2 \cup A_4$. Gambar 1 menunjukkan pembagian $\mathcal{X}$ dan Gambar 2 menunjukkan representasinya dalam sebuah pohon.



Gambar 1. Partisi $\mathcal{X}$ yang dibentuk oleh pemisahan ortogonal.



Gambar 2. Representasi pohon biner dari *classifier* yang sesuai dengan partisi.

2.2.2. *Model Trees*

Jenis pohon lain yang digunakan untuk prediksi numerik dikenal sebagai pohon model (*model trees*). Pohon model ditanam dengan cara yang sama seperti pohon regresi, tetapi pada setiap daun, model regresi linier berganda dibangun dari *example* yang mencapai simpulnya. Bergantung pada jumlah node daun, pohon model dapat membangun puluhan atau bahkan ratusan model. Hal ini menyebabkan pohon model lebih sulit dipahami daripada pohon regresi sejenisnya, namun dapat menghasilkan model yang lebih akurat [13].

Pohon yang dapat melakukan prediksi numerik memberikan alternatif yang menarik namun seringkali tidak dipertimbangkan untuk pemodelan regresi. Kelebihan

dan kekurangan pohon regresi dan pohon model relatif terhadap metode regresi yang lebih umum [13] tercantum dalam Tabel 2.

Tabel 2. Kelebihan dan kekurangan pohon regresi dan pohon model

Kelebihan	Kekurangan
• Memiliki keunggulan pohon keputusan dengan kemampuan untuk memodelkan data numerik • Pemilihan fitur otomatis • Tidak mengharuskan pengguna untuk menentukan model terlebih dahulu • Kemungkinan cocok dengan beberapa jenis data jika dibandingkan dengan regresi linier • Tidak memerlukan pengetahuan statistik untuk menginterpretasikan model	• Tidak umum digunakan (tidak seperti regresi linier yang lebih umum digunakan) • Membutuhkan sejumlah besar data pelatihan • Sulit untuk menentukan efek secara keseluruhan dari fitur individu berdasarkan hasil yang disajikan • Kemungkinan lebih sulit untuk ditafsirkan daripada model regresi

Metode regresi tradisional biasanya merupakan pilihan pertama untuk prediksi numerik, namun dalam beberapa kasus pohon keputusan numerik memberikan kelebihan yang lain. Misalnya, pohon keputusan mungkin lebih cocok untuk tugas-tugas dengan fitur yang banyak atau jika terdapat banyak hubungan non-linier yang kompleks antara fitur dan hasilnya. Pemodelan regresi juga membuat asumsi tentang bagaimana data numerik didistribusikan. Kriteria pemisahan yang umum disebut *standard deviation reduction* (SDR) ditunjukkan oleh persamaan (2).

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{2}$$

Pada persamaan (2), fungsi $sd(T)$ mengacu pada standar deviasi nilai di himpunan T, sedangkan T_i, T_2, ..., T_n adalah himpunan nilai yang dihasilkan dari pemisahan pada fitur. $|T|$ mengacu pada jumlah pengamatan pada himpunan T. Pada dasarnya, persamaan 2 mengukur pengurangan standar deviasi dari nilai asli ke standar deviasi tertimbang pasca pemisahan.

Misalnya, sebuah pohon memutuskan apakah akan melakukan pemisahan pada fitur biner A atau pemisahan pada fitur biner B seperti yang ditunjukkan oleh Gambar 3.

Data asli	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
Pemisahan pada fitur A	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
Pemisahan pada fitur B	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
				T1								**T2**			

Gambar 3. Pemisahan berdasarkan fitur

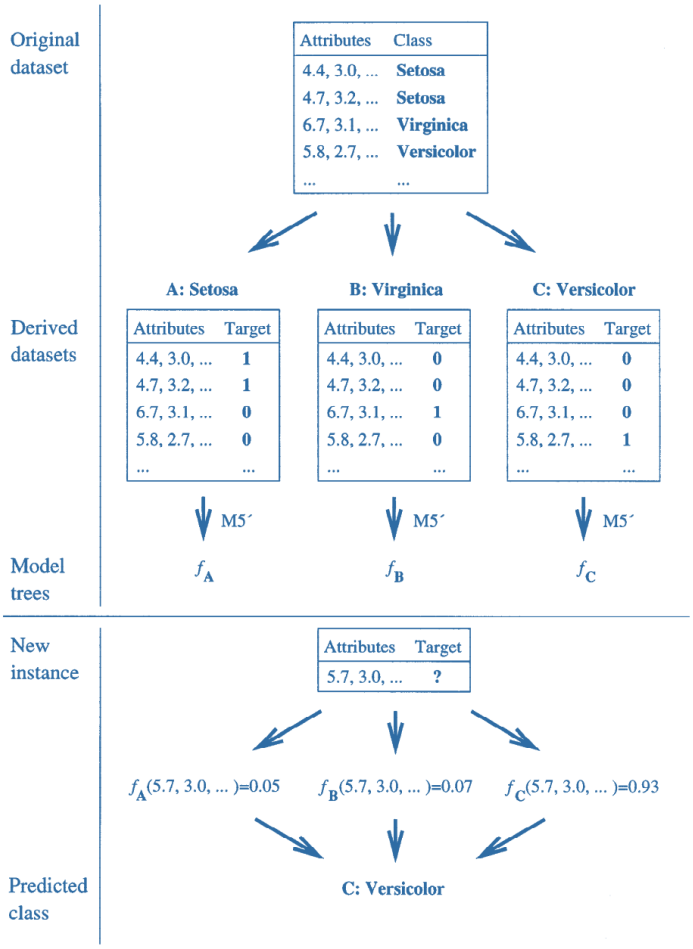
Pada studi ini pohon model yang digunakan direpresentasikan dalam algoritma M5'. Gambar 4 menunjukkan diagram bagaimana pembangun pohon model digunakan untuk klasifikasi dengan menggunakan dataset Iris. Pada Gambar 4, bagian atas menggambarkan proses pelatihan dan bagian bawah menggambarkan proses pengujian.

Pelatihan dimulai dengan menurunkan beberapa kumpulan data baru dari kumpulan data asli, yaitu satu untuk setiap kemungkinan nilai kelas. Dalam hal ini ada tiga dataset turunan, yaitu varietas Iris Setosa, Virginica dan Versicolor. Setiap dataset

turunan berisi jumlah *instance* yang sama dengan aslinya, dengan nilai kelas ditetapkan ke 1 atau 0 bergantung pada apakah *instance* tersebut memiliki kelas yang sesuai atau tidak. Pada langkah selanjutnya induser pohon model digunakan untuk menghasilkan pohon model pada setiap kumpulan data baru. Untuk *instance* tertentu, output dari satu pohon model merupakan perkiraan probabilitas bahwa *instance* tersebut milik kelas terkait. Karena nilai output dari pohon model hanyalah perkiraan, maka jumlahnya tidak harus satu.

Dalam proses pengujian, sebuah *instance* dari kelas yang tidak diketahui diproses oleh masing-masing pohon model, masing-masing hasilnya menjadi perkiraan probabilitas bahwa instance merupakan milik kelas terkait. Kelas yang pohon modelnya memberikan nilai tertinggi selanjutnya dipilih sebagai kelas prediksi.

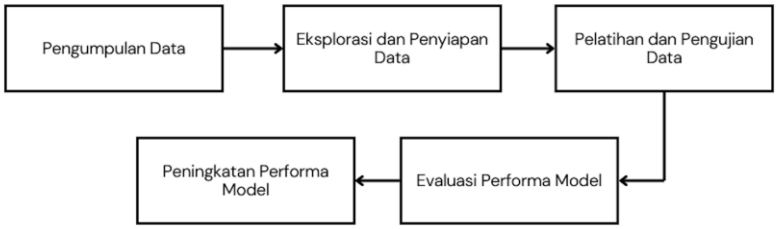
Prosedur pembelajaran M5' dilakukan dengan membagi ruang *instance* menjadi beberapa wilayah menggunakan pohon keputusan dan meminimalisir mean squared error yang diharapkan antara output pohon model dan nilai target 0 dan 1 untuk instance pelatihan dalam setiap wilayah tertentu. Instance pelatihan yang terletak pada wilayah tertentu dapat dilihat sebagai sampel dari distribusi probabilitas dasar yang menetapkan nilai kelas 0 dan 1 ke *instance* pada wilayah tersebut. Dalam statistik, prosedur standar ini digunakan untuk memperkirakan distribusi probabilitas dengan meminimalkan *mean squared error* dari sampel yang diambil [14]. Studi yang dilakukan oleh [15] menyimpulkan bahwa pohon model lebih akurat daripada pohon klasifikasi dan regresi untuk estimasi kedalaman gerusan.



Gambar 4. Klasifikasi menggunakan algoritma M5'

2.2.3. Metodologi Penelitian

Alur penelitian ditunjukkan oleh Gambar 4. Pengumpulan data dilakukan dengan menggunakan data *BodyFat* yang digunakan oleh Johnson [8]. Selanjutnya, data dieksplorasi dan disiapkan menggunakan perangkat lunak RStudio. Setelah melalui pra-pemrosesan, data dibagi menjadi data latih dan data uji. Selanjutnya, algoritma CART diterapkan pada data latih sehingga menghasilkan model untuk membuat prediksi. Setelah itu, model diterapkan pada data uji yang telah disiapkan sebelumnya. Evaluasi performa model diukur menggunakan *Mean Absolute Error* (MAE) yang dirumuskan dalam persamaan (3).



Gambar 5. Alur penelitian

Langkah selanjutnya adalah melakukan eksperimen peningkatan performa model menggunakan algoritma M5′ (M5-prime). Langkah terakhir adalah mengukur performa metode dengan menggunakan mekanisme yang sama seperti tahapan sebelumnya.

3. Hasil

3.1. Pengumpulan Data

Tahap pengumpulan data dilakukan dengan mengunduh Dataset BFP yang diusulkan oleh [8]. Dataset ini juga dapat diakses melalui laman https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset. Tabel 3 menunjukkan contoh potongan dataset BFP yang digunakan pada studi ini.

Tabel 3. Contoh potongan dataset

Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1,0708	12,3	23	154,25	67,75	36,2	93,1	85,2	94,5	59,0	37,3	21,9	32,0	27,4	17,1
1,0853	6,1	22	173,25	72,25	38,5	93,6	83,0	98,7	58,7	37,3	23,4	30,5	28,9	18,2
1,0414	25,3	22	154,00	66,25	34,0	95,8	87,9	99,2	59,6	38,9	24,0	28,8	25,2	16,6
1,0751	10,4	26	184,75	72,25	37,4	101,8	86,4	101,2	60,1	37,3	22,8	32,4	29,4	18,2
1,0340	28,7	24	184,25	71,25	34,4	97,3	100,0	101,9	63,2	42,2	24,0	32,2	27,7	17,7
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1,0271	31,9	74	207,50	70,00	40,8	112,4	108,5	107,1	59,3	42,2	24,6	33,7	30,0	20,9

Selanjutnya, data dibagi menjadi data latih dan data uji dengan rasio 3:1. Sejumlah 75% data digunakan sebagai data latih, yaitu data ke-1 sampai dengan data ke-189. Selanjutnya 25% sisanya digunakan sebagai data uji, yaitu data ke-190 sampai dengan data ke-252.

3.2. Eksplorasi dan Penyiapan Data

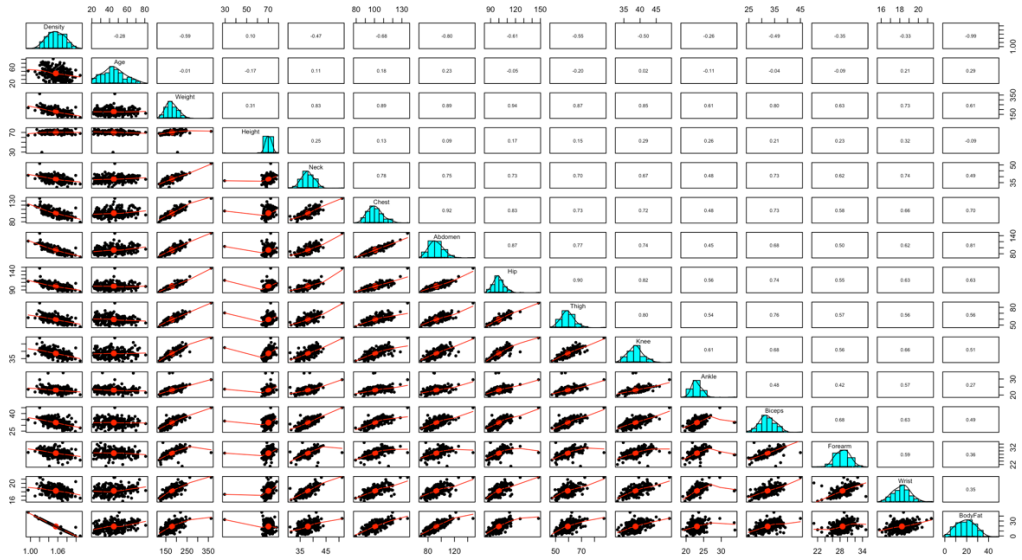
Tahap eksplorasi dan penyiapan data dilakukan dengan bantuan perangkat lunak RStudio. Gambar 5 menunjukkan 13 fitur yang terdapat pada dataset BFP dan 1 target prediksi, yaitu nilai BFP dari masing-masing perekaman data. Pada awalnya dataset BFP memiliki 14 fitur seperti yang ditunjukkan oleh Tabel 1. Namun pada penelitian ini,

variabel Density memiliki korelasi negatif dengan variabel BodyFat sehingga variabel Density akan dieliminasi.

```
> str(bodyfat2)
'data.frame':   252 obs. of  14 variables:
 $ BodyFat: num  12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...
 $ Age    : int  23 22 22 26 24 24 26 25 25 23 ...
 $ Weight : num  154 173 154 185 184 ...
 $ Height : num  67.8 72.2 66.2 72.2 71.2 ...
 $ Neck   : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
 $ Chest  : num  93.1 93.6 95.8 101.8 97.3 ...
 $ Abdomen: num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
 $ Hip    : num  94.5 98.7 99.2 101.2 101.9 ...
 $ Thigh  : num  59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...
 $ Knee   : num  37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...
 $ Ankle  : num  21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...
 $ Biceps : num  32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...
 $ Forearm: num  27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...
 $ Wrist  : num  17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...
```

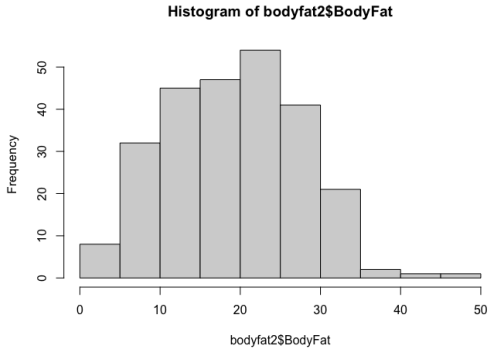
Gambar 6. Korelasi antar variabel pada dataset BFP

Ringkasan korelasi antar variabel pada dataset BFP dibuat menggunakan RStudio yang ditunjukkan oleh Gambar 7.



Gambar 7. Korelasi antar variabel pada dataset BFP

Uji normalitas data dilakukan dengan menggunakan Uji Shapiro-Wilk dengan menggunakan bantuan perangkat lunak RStudio. Gambar 7 menunjukkan histogram hasil Uji Shapiro-Wilk terhadap fitur target BodyFat pada dataset BFP.



Gambar 8. Histogram variabel target BodyFat

```
> shapiro.test(bodyfat2$BodyFat)

        Shapiro-Wilk normality test

data:  bodyfat2$BodyFat
W = 0.99168, p-value = 0.1649
```

Gambar 9. Hasil uji normalitas data

Gambar 8 menunjukkan hasil Uji Shapiro-Wilk di mana nilai p-value adalah 0,1649 (p-value > 0,05) sehingga data yang digunakan pada penelitian ini dapat dikatakan terdistribusi normal.

3.3. Pelatihan dan Pengujian Data

Pelatihan data diawali dengan membuat model pohon regresi, yaitu dengan menginstall paket **rpart** pada RStudio. Fungsi rpart() memuat pohon regresi menggunakan sintaks berikut:

- Sintaks pohon regresi: gunakan fungsi **rpart()** pada paket rpart
- Membuat model: **m <- rpart(dv ~ iv, data = mydata)**
  - dv adalah variabel tak bebas pada data frame mydata yang akan dimodelkan
  - iv adalah rumus R yang menentukan variabel bebas pada data frame mydata yang digunakan pada model
  - data merupakan data frame yang memuat variabel dv dan iv

  Fungsi tersebut akan menghasilkan objek model pohon regresi yang dapat digunakan untuk membuat prediksi.
- Membuat prediksi: **p <- predict(m, test, type = "vector")**
  - m adalah model yang dilatih oleh fungsi rpart()
  - test adalah data frame yang memuat data uji dengan fitur yang sama seperti pada data latih yang digunakan untuk membuat model
  - type menentukan jenis dari prediksi, dapat berupa "vector" (untuk prediksi nilai numerik), "class" untuk kelas yang diprediksi, atau "prob" (untuk kelas probabilitas yang diprediksi).

  Fungsi tersebut akan menghasilkan suatu vektor prediksi tergantung pada jenis parameter.

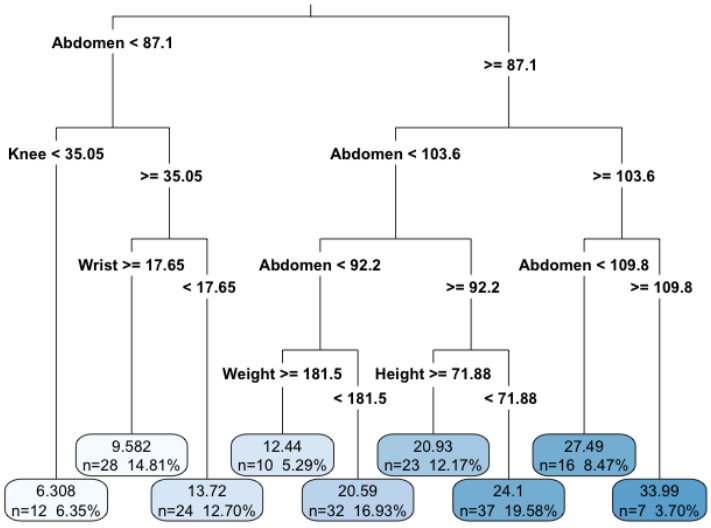
Setelah model pohon regresi diimplementasikan pada data frame BFP, maka model akan menghasilkan pohon keputusan yang secara sederhana ditunjukkan oleh Gambar 10.

```
> bodyfat2m.rpart
n= 189

node), split, n, deviance, yval
      * denotes terminal node

 1) root 189 12068.78000 18.558200
   2) Abdomen< 87.1 64  1417.04400 10.520310
     4) Knee< 35.05 12   168.00920  6.308333 *
     5) Knee>=35.05 52   987.01690 11.492310
      10) Wrist>=17.65 28   436.04110  9.582143 *
      11) Wrist< 17.65 24   329.61960 13.720830 *
   3) Abdomen>=87.1 125  4399.78300 22.673600
     6) Abdomen< 103.55 102  2508.74700 21.141180
      12) Abdomen< 92.2 42  1085.64500 18.652380
        24) Weight>=181.5 10   116.86400 12.440000 *
        25) Weight< 181.5 32   462.23880 20.593750 *
      13) Abdomen>=92.2 60   980.84330 22.883330
        26) Height>=71.875 23   254.28870 20.930430 *
        27) Height< 71.875 37   584.30970 24.097300 *
     7) Abdomen>=103.55 23   589.24870 29.469570
      14) Abdomen< 109.8 16   319.56940 27.493750 *
      15) Abdomen>=109.8 7    64.44857 33.985710 *
```

Gambar 10. Informasi dasar pada pohon keputusan dataset BFP

Untuk setiap simpul dalam pohon, jumlah sampel yang mencapai titik keputusan dicantumkan. Misalnya, seluruh data (sebanyak 252 data) dimulai dari simpul akar, di mana 64 data memiliki Abdomen < 87,1 dan 125 data memiliki Abdomen >= 87,1. Karena fitur Abdomen merupakan fitur yang pertama digunakan pada pohon, maka Abdomen menjadi satu-satunya prediktor presentase lemak tubuh yang paling penting. Node yang ditunjukkan dengan tanda * adalah terminal atau node daun, yang berarti menghasilkan prediksi (disebut sebagai *yval*). Sebagai contoh, simpul 4 memiliki *yval* 6,308333. Ketika pohon digunakan untuk melakukan prediksi, setiap sampel dengan Abdomen < 87,1 dan Knee < 35,05 akan diprediksi memiliki presentase lemak tubuh sebesar 6,31.

Selain dengan menggunakan informasi dasar seperti pada Gambar 10, pohon keputusan dataset BFP juga dapat ditinjau dengan menggunakan visualisasi diagram pohon yang ditunjukkan oleh Gambar 11.



Gambar 11. Visualisasi pohon model dataset BFP

3.4. Evaluasi Performa Model

Statistik ringkasan prediksi pada penelitian ini menunjukkan suatu potensi permasalahan, di mana prediksi jatuh pada kisaran yang jauh lebih sempit daripada nilai aktual seperti yang ditunjukkan oleh Tabel 4.

Tabel 4. Statistik ringkasan prediksi dibandingkan dengan nilai aktual

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Ringkasan Nilai Prediksi	6,308	13,721	20,930	20,884	24,097	33,986
Ringkasan Nilai Aktual	5,20	13,05	19,50	20,93	27,25	47,50

Temuan ini menunjukkan bahwa model tersebut tidak mengidentifikasi kasus ekstrim dengan benar, khususnya pada nilai persentase lemak tubuh yang terkecil dan terbesar. Korelasi antara nilai prediksi dan aktual menyediakan cara sederhana untuk mengukur performa model. Fungsi cor() pada RStudio dapat digunakan untuk mengukur hubungan antara dua vektor yang memiliki panjang sama, dengan deskripsi rumus yang ditunjukkan oleh persamaan (3).

$$\rho_x = Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} \quad (3)$$

di mana,
ρ_x = Korelasi Pearson
$Cov(x, y)$ = Kovarians x dan y
σ = Deviasi standar

Dengan menggunakan rumus korelasi, dapat diketahui bahwa nilai prediksi dan nilai aktual pada dataset BFP menggunakan algoritma CART memiliki korelasi sebesar 0,78 di mana nilai tersebut menunjukkan hubungan linier yang cukup kuat. Namun, nilai korelasi hanya mengukur seberapa kuat prediksi terkait dengan nilai aktualnya dan tidak mengukur seberapa jauh nilai prediksi dari nilai aktualnya.

Performa model juga diukur dengan mempertimbangkan seberapa jauh, rata-rata prediksinya dari nilai aktualnya. Pengukuran ini disebut *Mean Absolute Error* (MAE) seperti yang ditunjukkan oleh Persamaan (4), di mana n mengindikasikan jumlah prediksi, sedangkan e mendefinisikan jumlah kesalahan untuk prediksi i.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i| \quad (4)$$

Dengan menggunakan persamaan (4), nilai MAE untuk prediksi pada penelitian ini adalah 4,85. Nilai ini menyiratkan bahwa rata-rata perbedaan antara prediksi model dan nilai aktualnya adalah sekitar 4,85. Hal ini menunjukkan bahwa model bekerja dengan cukup baik mengingat berdasarkan skala, nilai *BodyFat* terkecil adalah 0 dan nilai *BodyFat* terbesar adalah 47,50.

3.5. Peningkatan Performa Model

Pada penelitian ini, pohon model digunakan untuk meningkatkan kinerja pembelajar. Pohon model meningkatkan pohon regresi dengan cara mengganti simpul daun dengan model regresi. Cara ini seringkali menghasilkan hasil yang lebih akurat daripada pohon regresi, yang hanya menggunakan satu nilai untuk prediksi pada simpul daun. Metode termutakhir dalam pohon model adalah algoritma M5' (M5 prime). Dengan menggunakan algoritme M5′, pohon regresi yang sudah ada (dengan menggunakan algoritme CART) dapat mengalami peningkatan performa karena simpul daun yang ada digantikan dengan model regresi. Peningkatan performa model menggunakan algoritma M5′ bertujuan untuk menaikkan nilai korelasi dan menurunkan nilai MAE. Algoritme M5′ tersedia pada R melalui paket RWeka dan fungsi M5P() yang memiliki sintaks sebagai berikut:

- Sintaks pohon model: menggunakan fungsi **M5P()** pada paket RWeka
- Membuat model: **m <- M5P(dv ~ iv, data = mydata)**
    - dv adalah variabel tak bebas pada data frame mydata yang akan dimodelkan
    - iv adalah rumus R yang menentukan variabel bebas pada data frame mydata yang digunakan pada model
    - data merupakan data frame yang memuat variabel dv dan iv

  Fungsi tersebut akan menghasilkan objek pohon model yang dapat digunakan untuk membuat prediksi.
- Membuat prediksi: **p <- predict(m, test)**
    - m adalah model yang dilatih oleh fungsi M5P()
    - test adalah data frame yang memuat data uji dengan fitur yang sama seperti pada data latih yang digunakan untuk membuat model

  Fungsi tersebut akan menghasilkan suatu vektor dari nilai numerik yang diprediksi.

Gambar 12 merupakan hasil dari implementasi algoritma M5′ pada dataset BFP:

```
> bodyfat2m.m5p
M5 pruned model tree:
(using smoothed linear models)

Abdomen <= 91.9 : LM1 (132/51.021%)
Abdomen >  91.9 : LM2 (120/48.468%)

LM num: 1
BodyFat =
        0.0055 * Age
        - 0.0093 * Weight
        - 0.3733 * Height
        - 0.6335 * Neck
        + 0.9303 * Abdomen
        - 0.0222 * Hip
        + 0.0277 * Thigh
        + 0.0305 * Biceps
        - 1.1586 * Wrist
        + 5.7281

LM num: 2
BodyFat =
        0.0059 * Age
        - 0.1086 * Weight
        - 0.0483 * Neck
        + 0.8325 * Abdomen
        - 0.0242 * Hip
        + 0.0302 * Thigh
        + 0.0332 * Biceps
        - 1.4611 * Wrist
        - 8.9983

Number of Rules : 2
```

Gambar 12. Implementasi algoritma M5′ pada dataset BFP

Sedangkan untuk evaluasi menggunakan algoritma M5′ didapatkan nilai korelasi sebesar 0,86 dan nilai MAE sebesar 3,86. Tabel 4 berikut menunjukkan perbandingan hasil korelasi dan MAE pada algoritma CART dan algoritma M5′.

Tabel 5. perbandingan hasil korelasi dan MAE pada algoritma CART dan algoritma M5′

Metode	Korelasi	MAE
CART	0,78	4,85
M5′	**0,86**	**3,86**

4. Pembahasan

Pada Tabel 5 dapat dilihat bahwa terdapat peningkatan nilai korelasi dari algoritma CART dan algoritma M5′ dan MAE yang mengalami penurunan. Proses pemisahan fitur pada algoritma M5′ sangat mirip dengan pohon regresi saat menggunakan algoritma CART. Abdomen merupakan variabel yang paling penting. Namun, node berakhir bukan dalam prediksi numerik, tetapi model linier (ditunjukkan sebagai LM1 dan LM2 pada Gambar 12).

```
LM num: 1
BodyFat =
        0.0055 * Age
        - 0.0093 * Weight
        - 0.3733 * Height
        - 0.6335 * Neck
        + 0.9303 * Abdomen
        - 0.0222 * Hip
        + 0.0277 * Thigh
        + 0.0305 * Biceps
        - 1.1586 * Wrist
        + 5.7281
```

Gambar 13. Implementasi algoritma M5′ pada dataset BFP

Model linier akan ditampilkan pada outoput. Misalnya, model untuk LM1 dideskripsikan pada Gambar 12. Nilai-nilai tersebut dapat diinterpretasikan persis sama dengan model regresi berganda. Setiap angka adalah efek bersih dari fitur terkait pada

nilai Body Fat yang diprediksi. Koefisien 0,93 untuk Abdomen menyiratkan bahwa untuk setiap peningkatan 1 unit Abdomen, Body Fat akan meningkat sebesar 0,93.

5. Kesimpulan

Pada penelitian ini telah disajikan metode pohon keputusan untuk prediksi nilai presentase lemak tubuh pada manusia. Kedua metode pohon keputusan yang digunakan tersebut meliputi pohon regresi (algoritma CART) dan pohon model (algoritma M5'). Algoritma CART menggunakan nilai rata-rata sampel pada simpul daun untuk membuat prediksi numerik, sedangkan algoritma M5' membangun model regresi pada setiap simpul daun dengan pendekatan hibrid. Pohon regresi memberikan cara sederhana untuk menjelaskan hubungan antara fitur dan hasil numerik, tetapi pohon model yang lebih kompleks juga memberikan hasil yang lebih akurat. Hasil menunjukkan algoritma M5' lebih unggul pada dataset BFP dengan nilai korelasi sebesar 0,86 dan nilai MAE sebesar 3,86.

Referensi

[1] World Health Organization, "Obesity and overweight," *Obesity and overweight*, Jun. 09, 2021. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed Oct. 04, 2022).

[2] Y. E. Shao, "Body Fat Percentage Prediction Using Intelligent Hybrid Approaches," *The Scientific World Journal*, vol. 2014, pp. 1–8, 2014, doi: 10.1155/2014/383910.

[3] A. Kupusinac, E. Stokić, and R. Doroslovački, "Predicting body fat percentage based on gender, age and BMI by using artificial neural networks," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 610–619, Feb. 2014, doi: 10.1016/j.cmpb.2013.10.013.

[4] M. Arroyo, A. M. Rocandio, L. Ansotegui, H. Herrera, I. Salces, and E. Rebato, "Comparison of predicted body fat percentage from anthropometric methods and from impedance in university students," *Br J Nutr*, vol. 92, no. 5, pp. 827–832, Nov. 2004, doi: 10.1079/BJN20041273.

[5] S. Meeuwsen, G. W. Horgan, and M. Elia, "The relationship between BMI and percent body fat, measured by bioelectrical impedance, in a large adult sample is curvilinear and influenced by age and sex," *Clinical Nutrition*, vol. 29, no. 5, pp. 560–566, Oct. 2010, doi: 10.1016/j.clnu.2009.12.011.

[6] Batech, M., Beeson, W.L., Schultz, E., Salto, L., Firek, A., DeLeon, M., Balcazar, H. and Cordero-MacIntyre, Z., "Comparison of Body Composition by Bioelectrical Impedance Analysis and Dual-Energy X-Ray Absorptiometry in Hispanic Diabetics".

[7] R. J. Maughan, "An evaluation of a bioelectrical impedance analyser for the estimation of body fat content.," *British Journal of Sports Medicine*, vol. 27, no. 1, pp. 63–66, Mar. 1993, doi: 10.1136/bjsm.27.1.63.

[8] R. W. Johnson, "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, vol. 4, no. 1, p. 6, Mar. 1996, doi: 10.1080/10691898.1996.11910505.

[9] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," *Applied Soft Computing*, vol. 14, pp. 47–52, Jan. 2014, doi: 10.1016/j.asoc.2013.09.020.

[10] K. W. DeGregory *et al.*, "A review of machine learning in obesity: Machine learning in obesity research," *Obesity Reviews*, vol. 19, no. 5, pp. 668–685, May 2018, doi: 10.1111/obr.12667.

[11] L. A. R. Hakim, A. A. Rizal, and D. Ratnasari, "Aplikasi Prediksi Kelulusan Mahasiswa Berbasis K-Nearest Neighbor (K-NN)," *jtim*, vol. 1, no. 1, pp. 30–36, May 2019, doi: 10.35746/jtim.v1i1.11.

[12] L. Breimann, J. Friedman, C. Stone, and R. Olshen, "Classification and regression trees, 25 Chapman & Hall," *CRC, Wadsworth, Belmont, California*, 1984.

[13] B. Lantz, "Machine Learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real world applications," in *Livery Place*, Packt Publishing Ltd., Packt Publishing Ltd, 2013.

[14] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. New York, NY: Springer New York, 1996. doi: 10.1007/978-1-4612-0711-5.

[15] M. Samadi, E. Jabbari, and H. Md. Azamathulla, "Assessment of M5' model tree and classification and regression trees for prediction of scour depth below free overfall spillways," *Neural Comput & Applic*, vol. 24, no. 2, pp. 357–366, Feb. 2014, doi: 10.1007/s00521-012-1230-9.