JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

Pemodelan Classification and Regression Tree (CART) Pada Klasifikasi Gaya Hidup Sehat Menggunakan Pendekatan User-Based Classification

Anita Sindar Sinaga[1], Bella Saputri[2], Nadia Aulia[3]
[1,2,3] Program Studi Teknologi Informasi, STMIK Pelita Nusantara
Email: [1,*]haito_ita@yahoo.com, [2]bella@gmail.com, [3]nadia@gmail.com
Email Penulis Korespondensi: haito_ita@yahoo.com

Abstrak
Penentuan gaya hidup sehat menjadi isu penting dalam bidang kesehatan masyarakat, terutama dalam upaya pencegahan penyakit kronis. Proses klasifikasi dilakukan dengan menyusun pohon keputusan yang membagi data ke dalam kelas gaya hidup (sehat/tidak sehat) secara rekursif, berdasarkan fitur yang memberikan pemisahan terbaik. Hasil penelitian menunjukkan bahwa model CART mampu mengidentifikasi pola-pola gaya hidup yang signifikan dengan akurasi klasifikasi yang cukup tinggi, serta memberikan pemahaman yang jelas tentang faktor-faktor pengguna yang berkontribusi terhadap status gaya hidup sehat. Dari total 70 data, nilai target rata-ratanya adalah 0.146. Tree akan membagi data berdasarkan nilai Feature ≤ 3.25. Jika benar (True) ke kiri, jika salah (False) ke kanan. Node ini menunjukkan 49 sampel dari root mengikuti kondisi Feature <= 3.25 dan sekarang dibagi lagi berdasarkan Feature <= 2.538. Proses ini berlanjut terus secara rekursif hingga mencapai leaf node. Ada 1 sampel, dengan nilai target = 0.3, sehingga tidak ada variansi (squared_error = 0). Jika data baru masuk ke cabang ini, model akan memprediksi 0.3 sebagai nilai regresi.
Kata Kunci: Klasifikasi, Gaya Hidup Sehat, Regresi, CART, User-Based

Abstract
Determining a healthy lifestyle is an important issue in public health, especially in efforts to prevent chronic diseases. The classification process is carried out by constructing a decision tree that divides data into lifestyle classes (healthy/unhealthy) recursively, based on the features that provide the best separation. The results show that the CART model is able to identify significant lifestyle patterns with fairly high classification accuracy, as well as provide a clear understanding of user factors that contribute to healthy lifestyle status. This approach supports decision making in data-based health promotion programs. From a total of 70 data, the average target value is 0.146. The tree will divide the data based on whether the Feature value is ≤ 3.25. If true to the left, if false to the right. This node shows 49 samples from the root following the condition Feature <= 3.25 and is now divided again based on Feature <= 2.538. This process continues recursively until it reaches the leaf node. There is only 1 sample, with target value = 0.3, so there is no variance (squared_error = 0). If new data enters this branch, the model will predict 0.3 as the regression value.
***Keywords**: Classification, Healthy Lifestyle, Regression, CART, User-Based Classification*

1. PENDAHULUAN

Hasil survei yang diselenggarakan Kementerian Kesehatan, diperoleh lebih dari 50%, sekarang masyarakat Indonesia, memiliki kebiasaan hidup dengan giat melaksanakan pola kehidupan yang seimbang, melakukan kegiatan olahraga, menjaga kesehatan mental dan melakukan meditasi. Menerapkan pola hidup yang seimbang, mengubah kualitas hidup menjadi semakin membaik dan dapat terhindar dari serangan penyakit [1]. Menurut WHO, membiasakan hidup sehat dengan menjaga keadaan kesejahteraan fisik, mental, dan sosial secara menyeluruh, tidak hanya terbebas dari serangan penyakit atau mudah mengalami paparan virus penyakit. WHO mengingatkan pentingnya mempunyai cara hidup sehat dan bersih (PHBS) sebagai bagian terutama dari gaya hidup sehat. Faktor yang mempengaruhi kualitas keadaan hidup seseorang beragam, dipengaruhi faktor internal (dari dalam diri individu) dan faktor eksternal (dari lingkungan sekitar). Kondisi hidup yang kurang sehat dapat memperbesar kemungkinan terjadinya penyakit kronis seperti diabetes, gangguan jantung, obesitas, serta gangguan pada sistem pernapasan.

Analisisis klasifikasi dipergunakan untuk melihat hubungan antara beberapa variabel prediktor dengan satu variabel respon yang bersifat kategori, variabel prediktor digunakan untuk menentukan kelas dari variabel respon. Variabel-variabel predictor dimanfaatkan untuk memperkirakan kategori atau kelas dari variabel respon [2]. Pendekatan ini bertujuan untuk mengidentifikasi keterkaitan antara variabel-variabel prediktor dengan satu variabel respon kategorikal, serta memanfaatkan variabel-variabel tersebut untuk mengklasifikasikan nilai dari respon. Perancangan model pengklasifikasian dengan analsisis klasifikasi dapat mengkategorikan data ke dalam kelas-kelas berbeda berdasarkan kriteria tertentu. Teknik klasifikasi terhadap data baru dilakukan dengan memanfaatkan dan mengolah data yang sebelumnya telah diklasifikasi, kemudian menggunakan hasil analisis tersebut untuk membentuk sejumlah aturan keputusan. Klasifikasi data baru dilakukan dengan mengolah data lama yang sudah dikategorikan, lalu menghasilkan aturan-aturan tertentu yang dapat diterapkan pada data baru [3]. Model dilatih menggunakan data training, lalu dievaluasi pada data uji sebelum digunakan untuk melakukan prediksi pada data baru yang belum terlihat. Berdasarkan data-data yang bersumber dari pola makan sehat, aktivitas fisik yang teratur, istirahat yang cukup, pengelolaan stres, serta menghindari kebiasaan buruk seperti merokok dan konsumsi alkohol berlebihan maka dapat dikelompokkan seseorang

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

menjalankan gaya hidup sehat atau tidak. Klasifikasi ini membantu mengelompokkan gaya hidup sehat menjadi aspek yang spesifik dan terukur, sehingga memudahkan desain program, pengembangan indikator, dan evaluasi pemodelan klasifikasi [4].

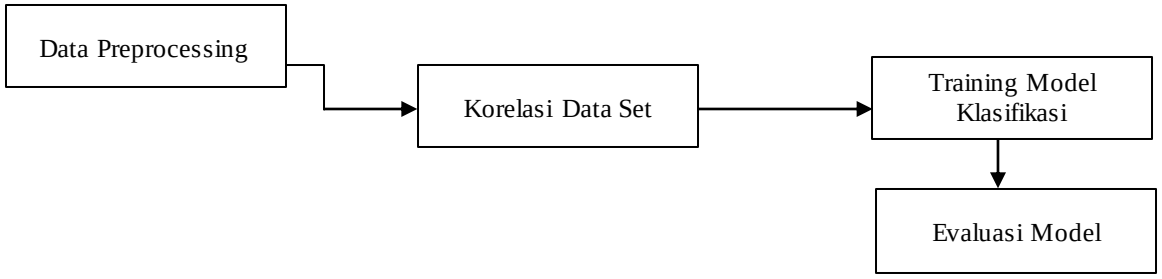
Metode klasifikasi dalam Machine Learning atau data mining menggunakan atribut atau profil pengguna untuk mengklasifikasikan atau memprediksi suatu label atau outcome[5]. *User-Based Classification* berfokus pada peran dan pemahaman pengguna dalam proses pembuatan, pengeditan, evaluasi, maupun distribusi data. Proses klasifikasi memanfaatkan algoritma untuk memprediksi kelas atau kategori dari suatu data berdasarkan nilai-nilai yang terdapat dalam data [6]. Beragam algoritma dapat digunakan untuk keperluan klasifikasi, tergantung pada jenis data dan kompleksitas masalah. Beberapa algoritma yang umum digunakan dalam klasifikasi antara lain adalah *Logistic Regression, Decision Tree, Random Forest, Support Vector Machine* (SVM), hingga metode berbasis *neural network*. Pendekatan user-based, klasifikasi menjadi lebih kontekstual dan relevan, karena mempertimbangkan data-data yang merepresentasikan perilaku, karakteristik, atau kebiasaan pengguna secara langsung. Dengan menggunakan algoritma seperti *Random Forest, SVM,* atau *Logistic Regression*, membangun *User-Based Classifier* yang memetakan karakteristik gaya hidup sehat ke dalam kelas tertentu *User-Based Classification* merupakan klasifikasi yang berpusat pada data pengguna dapat digunakan untuk prediksi personal, segmentasi, rekomendasi, atau diagnosis [7].

Algoritma *Random Forest* bekerja dengan menumbuhkan banyak pohon keputusan *(trees)* dan memperkenalkan unsur keacakan dalam prosesnya. Tidak seperti algoritma konvensional yang selalu memilih fitur terbaik untuk pemisahan node, *Random Forest* memilih fitur terbaik dari subset fitur yang dipilih secara acak, sehingga meningkatkan keragaman model dan mengurangi risiko *overfitting* [8]. Selain itu, model Classification and Regression Tree (CART) juga digunakan untuk mengklasifikasikan komponen gaya hidup sehat berdasarkan deskripsi perilaku individu. CART dirancang untuk menganalisis variabel respon yang bersifat nominal, ordinal, maupun kontinu [9]. Keunggulan utama dari CART terletak pada kemampuannya dalam menyeleksi variabel-variabel prediktor yang paling berpengaruh serta mengidentifikasi interaksi antar variabel yang signifikan terhadap hasil klasifikasi [10]. Diperlukan sistem klasifikasi yang dapat mengidentifikasi kategori gaya hidup atau tingkat risiko kesehatan berdasarkan karakteristik pengguna secara individual. Penelitian ini membangun model klasifikasi gaya hidup sehat berdasarkan karakteristik penggun a *(user-based)* bertujuan membangun model klasifikasi berbasis pengguna untuk mengkategorikan gaya hidup sehat dan mengidentifikasi variabel gaya hidup yang paling memengaruhi kategori gaya hidup. .

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini termasuk dalam jenis penelitian kuantitatif dengan pendekatan deskriptif dan prediktif. Tahapan yang akan diterapkan pada penelitian untuk mencapai tujuan dari penelitian.



Gambar 1. Tahapan Penelitian

Uraian tahapan penelitian :

1. Data Preprocessing

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa dataset sintetik yang berisi variabel-variabel gaya hidup, seperti umur, jenis kelamin, indeks massa tubuh (BMI), pola makan, aktivitas fisik, stres, tidur, konsumsi alkohol, dan kebiasaan merokok. Variabel respon dalam penelitian ini adalah status kesehatan pengguna terkait risiko penyakit kronis (sehat atau berisiko). Data dikumpulkan melalui pengolahan dataset yang tersedia dalam bentuk file CSV. Proses ini mencakup pengolahan data awal, pembersihan data, dan transformasi data kategorikal menjadi numerik menggunakan teknik label encoding. Pra-pemrosesan seperti menangani nilai yang hilang, mengkodekan variabel kategorikal (jika ada), dan menskalakan fitur numerik jika diperlukan (meskipun Decision Tree kurang sensitif terhadap penskalaan dibandingkan beberapa algoritma lain). Bagi dataset menjadi set pelatihan (training set) dan set pengujian (testing set).

2. Korelasi Data Set

Data set menggambarkan nilai untuk setiap variabel untuk kuantitas yang tidak diketahui. Dua set data dikatakan berkorelasi jika keduanya tidak independen satu sama lain. Variabel yang menunjukkan hubungan linier antara satu sama

lain. Korelasi menunjukkan apakah ada hubungan prediktif. Korelasi data set mengacu pada hubungan atau keterkaitan antara dua atau lebih variabel dalam suatu kumpulan data. Jika dua variabel berkorelasi, bahwa satu variabel menyebabkan perubahan pada variabel lain. Korelasi hanya menilai hubungan antara variabel, dan mungkin ada faktor-faktor berbeda yang menyebabkan hubungan tersebut. Pemilihan fitur yang tepat dipengaruhi oleh variabel target. Fitur-fitur yang dimasukkan ke dalam model perlu memiliki keterkaitan yang signifikan dengan variabel target agar model mampu menghasilkan prediksi yang akurat.

3. Training Model Klasifikasi

Pemodelan prediktif dalam klasifikasi berfokus pada pengelompokan data ke dalam kategori yang telah ditentukan berdasarkan nilai fitur. Model machine learning jenis ini membagi data ke dalam kelas-kelas yang telah dikenal sebelumnya. Proses ini dibangun dari data historis untuk membentuk model yang dapat mengklasifikasikan data baru dengan tingkat akurasi yang dapat diterima. Dalam klasifikasi, variabel target bersifat diskrit, sehingga model dilatih untuk menghasilkan output diskrit (y) berdasarkan input tertentu (x), yang hasilnya berupa label atau kelas tertentu. Sebagai bagian dari pembelajaran mesin terawasi *(supervised learning)*, model ini dirancang untuk memprediksi label yang benar dari data masukan, dengan pelatihan dilakukan sepenuhnya pada dataset pelatihan dan kemudian diuji pada dataset pengujian sebelum digunakan untuk memprediksi data yang belum pernah dilihat sebelumnya. Model klasifikasi ini memungkinkan pengkategorian data berdasarkan atribut tertentu, sehingga mendukung prediksi yang lebih akurat dan memperkaya proses analisis data.

4. Evaluasi Model

Dalam data mining, text mining ataupun *Machine Learning*, evaluasi model juga penting untuk memastikan bahwa model bekerja dengan baik dalam menghasilkan prediksi yang tepat. Evaluasi model memastikan kualitas dan keandalan model. Target variable juga digunakan untuk mengevaluasi kinerja model. Dalam model klasifikasi, menggunakan metrik seperti akurasi, *precision*, *recall*, dan *F1-score* untuk menilai seberapa baik model memprediksi target variable. Dengan menggunakan metode evaluasi yang tepat, mengidentifikasi kekuatan dan kelemahan model, serta melakukan perbaikan yang diperlukan untuk mencapai tujuan yang diinginkan. Evaluasi kinerja model pada set pengujian menggunakan metrik yang sesuai dengan tugas (akurasi, *presisi*, *recall*, *F1-score* untuk klasifikasi *Mean Squared Error*, *R-squared* untuk regresi). Visualisasikan pohon keputusan atau lihat pentingnya fitur *(feature importance)* untuk memahami variabel yang paling berpengaruh dalam prediksi.

2.2 User-Based Classification

Model klasifikasi termasuk dalam kategori supervised learning, di mana proses pelatihan dilakukan dengan menggunakan data yang telah diberi label [11]. Model kemudian dievaluasi menggunakan data uji untuk memastikan tingkat akurasi dan kemampuan generalisasi sebelum diterapkan pada data yang belum pernah dilihat sebelumnya. Ciri utama dari model klasifikasi adalah variabel keluarannya bersifat diskrit, sehingga output yang dihasilkan berupa label atau kategori tertentu. Salah satu pendekatan yang digunakan dalam model klasifikasi adalah *User-Based Classification*, yaitu metode klasifikasi yang berfokus pada atribut, perilaku, serta karakteristik pengguna [12]. Pendekatan ini membangun model berdasarkan data yang secara langsung berkaitan dengan individu pengguna, seperti usia, jenis kelamin, gaya hidup (misalnya pola makan, frekuensi olahraga, jam tidur, tingkat stres), serta riwayat kesehatan. Dengan memanfaatkan informasi personal tersebut, model dapat mengidentifikasi pola spesifik yang berguna dalam menentukan kelas atau kategori dari suatu data, sehingga meningkatkan akurasi dalam proses klasifikasi [13].

2.3 *Classification and Regression Tree* (CART)

Classification and Regression Tree (CART) menghasilkan pohon keputusan yang setiap simpulnya hanya memiliki dua cabang (pohon biner), berbeda dengan beberapa metode pohon keputusan lain yang bisa memiliki lebih dari dua cabang [14]. CART bekerja dengan membagi dataset secara rekursif berdasarkan nilai variabel independen yang paling efektif memisahkan data menjadi kelompok-kelompok yang lebih homogen. Algoritma CART menggunakan berbagai kriteria untuk menentukan bagaimana membagi data pada setiap simpul, seperti *Gini impurity* atau *entropy* untuk klasifikasi, dan varians untuk regresi [15]. Pohon keputusan yang dihasilkan CART relatif mudah diinterpretasikan, memungkinkan pengguna untuk memahami keputusan dibuat. CART menganalisis data yang dapat digunakan untuk berbagai tujuan, mulai dari klasifikasi sederhana hingga pemodelan prediktif yang kompleks. Metode *Classification and Regression Trees* (CART) menggunakan pendekatan penyekatan secara rekursif dan biner *(binary recursive partitioning)* dalam proses klasifikasinya. Jika variabel respon yang dianalisis berskala kategorik, maka CART akan membentuk struktur berupa pohon klasifikasi [16]. Sebaliknya, apabila variabel respon bersifat kontinu, algoritma ini akan membentuk pohon regresi[17]. CART bekerja dengan membagi data secara berulang ke dalam dua subset berdasarkan fitur yang memberikan pemisahan terbaik. Tujuan utama dari proses ini adalah untuk membentuk node yang homogen, yaitu node yang sebisa mungkin hanya mengandung data dari satu kelas [18]. Proses pembagian dilakukan sampai diperoleh struktur pohon yang optimal untuk memetakan hubungan antara fitur dan target.

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

3. HASIL DAN PEMBAHASAN

Classification and Regression Tree (CART) dapat digunakan membangun *Decision Tree* untuk klasifikasi, jika variabel target bersifat kategorikal dan Regresi, jika variabel target bersifat numerik/kontinu. Variabel kategorikal dikelompokkan menjadi beberapa kategori yang jumlahnya relatif sedikit. Target variable (atau variabel target) sebagai dependent variable atau variabel tergantung diperoleh dari model berdasarkan fitur-fitur input yang tersedia. Fitur-fitur dari data yang paling relevan dan informatif untuk proses klasifikasi dikumpulkan dari sumber data. Pengambilan keputusan dalam klasifikasi gaya hidup sehat yang berbasis data. *Decision Tree* untuk regresi, yaitu model yang memprediksi nilai numerik (target) berdasarkan nilai fitur masukan (feature), bukan klasifikasi kelas. Fitur ini akan digunakan untuk membedakan antar kelas.

Atribut atau fitur (features) : Fitur (Features) atau Variabel dalam penelitian ini yaitu Age, Gender, Height_cm, Weight_kg, BMI, Smoker, Exercise_Freq, Diet_Quality, Alcohol_Consumption, Stress_Level, Sleep_Hours.

label atau target variabel : Chronic_Disease

Tahapan awal pemodelan Machine Learning untuk menemukan format yang sesuai untuk pelatihan model dilakukan import dan eksplorasi data, pembersihan data (cleaning), outliers, duplikasi, encoding variabel kategorikal, normalisasi/skala data dan pemisahan fitur dan target dan split data training dan testing.

Tabel 1. Jenis Variabel Data Set

Nama Kolom	Jenis Variabel	Keterangan
Age	Numerik Kontinu	Usia pengguna
Gender	Kategorikal	Jenis kelamin (misal: Male/Female)
Height_cm	Numerik Kontinu	Tinggi badan dalam cm
Weight_kg	Numerik Kontinu	Berat badan dalam kg
BMI	Numerik Kontinu	Indeks massa tubuh (hasil perhitungan)
Smoker	Kategorikal	Perokok atau tidak (Yes/No)
Exercise_Freq	Ordinal Kategorikal	Frekuensi olahraga (misal: Never, Sometimes, Often)
Diet_Quality	Ordinal Kategorikal	Kualitas diet (misal: Poor, Average, Good)
Alcohol_Consumption	Ordinal Kategorikal	Konsumsi alkohol (Low, Medium, High)
Chronic_Disease	Kategorikal (Target)	Target (misalnya: Yes/No atau 1/0)
Stress_Level	Ordinal Kategorikal	Tingkat stres (Low, Medium, High)
Sleep_Hours	Numerik Kontinu	Rata-rata jam tidur per hari

Variable bernilai numerik yang kontinu dapat menyelesaikan masalah regresi. Setiap fitur *(variabel prediktor)*, CART mengevaluasi semua kemungkinan titik pemisah. Selanjutnya dihitung impurity (ketidakhomogenan) dari setiap pembagian. CART menggunakan Gini Impurity untuk mengukur seberapa bersih node:

$$gini\ (t) = 1 - \sum_{i=1}^{n} pi^2 \qquad (1)$$

Pi^2 = proporsi data kelas ke-i di node. Semakin rendah nilai Gini, semakin homogen data dalam node. Dipilih fitur dan nilai pemisah yang menghasilkan penurunan Gini terbesar (impurity paling kecil). Data dibagi ke dalam dua cabang (left dan right node). Misalnya mengklasifikasikan seseorang memiliki gaya hidup sehat berdasarkan variable BMI dan jumlah tidur per hari. Berdasarkan tipe data yang akan diprediksi, pohon keputusan untuk pemodelan klasifikasi umumnya disebut dengan pohon klasifikasi. Pemodelan regresi disebut pohon regresi. Dalam pohon klasifikasi, setiap node menunjukkan tes kondisi pada fitur, dan setiap cabang menunjukkan hasil dari tes tersebut. *Leaf node* atau daun pohon menunjukkan prediksi target. Pohon klasifikasi dalam metode pembelajaran mesin yang mudah dipahami dan diterapkan, karena visualisasi pohonnya mempermudah pemahaman dan interpretasi model.

CART menghasilkan aturan :
Jika BMI > 27.5:
  Jika Tidur < 6 jam → Tidak Sehat
  Jika Tidur ≥ 6 jam → Tidak Sehat
Jika BMI ≤ 27.5:
  Jika Tidur < 5 jam → Tidak Sehat
  Jika Tidur ≥ 5 jam → Sehat

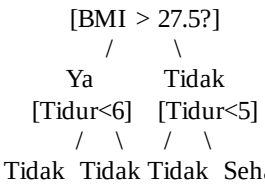
Fitur atau karakteristik dari suatu data yang digunakan dalam pohon klasifikasi untuk membuat pemisahan dan memprediksi label akhir. Pada setiap node dalam pohon klasifikasi, salah satu atribut dipilih untuk memisahkan data

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

menjadi subkelompok yang lebih kecil lagi. Aturan atau prosedur yang digunakan untuk memprediksi label akhir suatu data berdasarkan fitur atau atribut data tersebut. *Decision rules* dapat diambil dari setiap cabang dalam pohon klasifikasi dan merupakan pernyataan logika yang menentukan apakah data akan diklasifikasikan ke dalam satu label atau label lain. metodologi CART terdiri dari beberapa tahap utama, yaitu: pembangunan pohon secara penuh (maksimum), pemangkasan atau pemilihan ukuran pohon yang sesuai untuk menghindari *overfitting*, dan penggunaan pohon yang telah terbentuk untuk mengklasifikasikan data baru. Pendekatan ini sangat efektif dalam menangani baik masalah klasifikasi maupun regresi tergantung pada tipe data target yang digunakan. Rekursi diulang untuk masing-masing subtree :

1. Node hanya berisi satu kelas.
2. Jumlah data dalam node terlalu kecil.
3. Pohon mencapai kedalaman maksimum *(max depth)*.

Jika data tidak bisa dibagi lagi, terbentuk *leaf node* yang memuat kelas mayoritas dari data di dalamnya → itulah hasil klasifikasinya.



Gambar 2. Diagram Cabang Pohon Keputusan

3.1 Hasil

Pembentukan pohon klasifikasi diawali dengan menentukan variabel dan threshold untuk dijadikan pemilah tiap simpul. Masing-masing simpul dalam binary tree terdiri dari tiga bagian yaitu sebuah data dan dua buah pointer yang dinamakan pointer kiri dan kanan. Tiga elemen dalam satu *decision tree root node* (akar), *branches* (ranting), dan *leaf node* (daun), kemungkinan hasil atas setiap tindakan korelasi data menunjukkan seberapa kuat dan arah hubungan antara dua variabel numerik. Perubahan pada satu variabel cenderung diikuti oleh perubahan yang konsisten pada variabel lain.

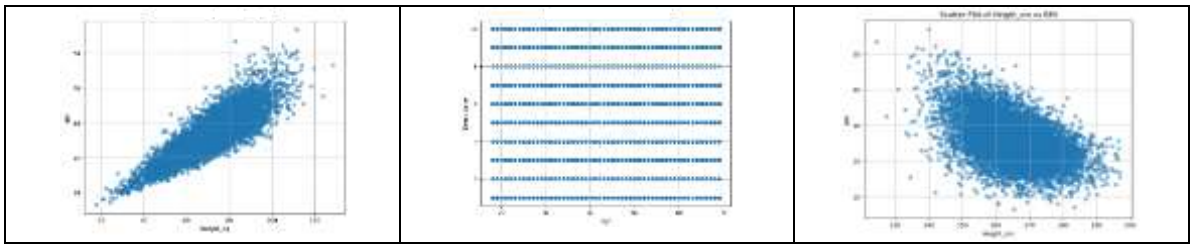
Tabel 2. Korelasi Data

	ID	Age	Height_cm	Weight_kg	BMI	Stress_Level	Sleep_Hours
ID	1.000000	-0.000441	0.015541	0.026077	0.013829	0.000932	-0.003942
Age	-0.000441	1.000000	-0.012390	-0.003737	0.004727	0.013866	-0.004060
Height_cm	0.015541	-0.012390	1.000000	-0.004695	-0.500452	0.008503	0.017716
Weight_kg	0.026077	-0.003737	-0.004695	1.000000	0.859116	-0.010306	-0.009961
BMI	0.013829	0.004727	-0.500452	0.859116	1.000000	-0.011619	-0.013451
Stress_Level	0.000932	0.013866	0.008503	-0.010306	-0.011619	1.000000	0.001539
Sleep_Hours	-0.003942	-0.004060	0.017716	-0.009961	-0.013451	0.001539	1.000000

Dari matriks korelasi diperoleh Weight_kg dan BMI memiliki korelasi positif yang kuat (0.86), (Height_cm) dan BMI memiliki korelasi negatif sedang (-0.50). Berikutnya menampilkan korelasi antara variabel kategorikal dan numerik, dan melakukan analisis univariat/bivariat. Diperoleh variabel kategorikal dan numerik dalam dataset:
['ID', 'Age', 'Height_cm', 'Weight_kg', 'BMI', 'Stress_Level', 'Sleep_Hours']
Categorical columns: ['Gender', 'Smoker', 'Exercise_Freq', 'Diet_Quality', 'Alcohol_Consumption', 'Chronic_Disease']
Dataset Variabel Numerik mencakup variabel numerik seperti Usia, Tinggi (cm), Berat (kg), BMI, Tingkat Stres, dan Jam Tidur. Analisis univariat menunjukkan distribusi yang bervariasi untuk setiap variabel, dengan beberapa variabel menunjukkan distribusi yang mendekati normal sementara yang lain mungkin miring atau memiliki nilai ekstrem (outlier). Variabel kategorikal meliputi Jenis Kelamin, Perokok, Frekuensi Olahraga, Kualitas Diet, Konsumsi Alkohol, dan Penyakit Kronis. Distribusi frekuensi untuk variabel-variabel ini bervariasi, menunjukkan proporsi yang berbeda dalam setiap kategori. Terdapat korelasi positif yang kuat antara Berat (kg) dan BMI (koefisien korelasi Pearson sekitar 0.8591). Ada korelasi negatif moderat antara Tinggi (cm) dan BMI (koefisien korelasi Pearson sekitar -0.5005). Tidak

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

ada korelasi yang signifikan antara Usia dan Tingkat Stres (koefisien korelasi sekitar 0.0139). Tidak ada korelasi yang signifikan antara Jam Tidur dan Tingkat Stres (koefisien korelasi sekitar 0.0015).



Gambar 3. Hubungan Antara Dua Variabel Numerik

Jika dua variabel memiliki korelasi positif yang kuat (mendekati +1), ini berarti ketika satu variabel meningkat, variabel lainnya juga cenderung meningkat. Ini menunjukkan bahwa variabel pertama bisa menjadi prediktor yang baik untuk variabel kedua. Contoh dari analisis sebelumnya adalah korelasi antara Weight_kg dan BMI. Berat badan yang lebih tinggi cenderung memprediksi BMI yang lebih tinggi. Jika dua variabel memiliki korelasi negatif yang kuat (mendekati -1), ini berarti ketika satu variabel meningkat, variabel lainnya cenderung menurun. Ini juga menunjukkan potensi hubungan prediktif. Contoh dari analisis sebelumnya adalah korelasi antara Height_cm dan BMI. Tinggi badan yang lebih tinggi cenderung memprediksi BMI yang lebih rendah (untuk berat yang sama). Korelasi Lemah atau Tidak Ada: Jika korelasi mendekati 0, ini berarti tidak ada hubungan linier yang kuat antara dua variabel. Dalam kasus ini, satu variabel mungkin bukan prediktor yang baik untuk variabel lainnya dalam model linier sederhana. Contoh dari analisis sebelumnya adalah korelasi antara Age dan Stress_Level, serta Sleep_Hours dan Stress_Level, yang sangat lemah.
Inisialisasi dan latih model Decision Tree (menggunakan DecisionTreeClassifier untuk klasifikasi atau DecisionTreeRegressor untuk regresi dari scikit-learn) pada set pelatihan.
Inisialisasi dan latih model Decision Tree (menggunakan *DecisionTreeClassifier* untuk klasifikasi atau DecisionTreeRegressor untuk regresi dari scikit-learn) pada set pelatihan. Target variable menentukan fokus dari model data analysis. Semua fitur atau variabel input dikumpulkan dan diproses dengan tujuan untuk memprediksi target variable. Tanpa target variable yang jelas, model tidak akan memiliki arah atau tujuan. Model klasifikasi dalam Machine Learning membagi titik data ke dalam kelompok yang telah ditentukan sebelumnya yang disebut kelas. Klasifikasi supervised learning bertujuan untuk memetakan input data ke dalam kelas tertentu berdasarkan data berlabel. Random forest menghasilkan beberapa pohon keputusan, dengan memilih fitur secara acak untuk membuat keputusan saat membagi node guna membuat setiap pohon mengambil pengamatan acak dari setiap pohon dan menghitung rata-ratanya untuk membangun model akhir. Klasifikasi CART membagi data ke dalam dua cabang berdasarkan fitur yang memberikan pemisahan terbaik. CART membuat pohon biner, setiap node hanya memiliki dua cabang. Model akan terus membagi data sampai mencapai kriteria tertentu.

3.2 Implementasi

Penelitian ini bertujuan untuk membangun model klasifikasi gaya hidup sehat menggunakan algoritma *Classification* and *Regression Tree* (CART) dengan pendekatan User-Based Classification. Model dikembangkan berdasarkan data pengguna yang mencakup atribut seperti usia, jenis kelamin, indeks massa tubuh (BMI), kebiasaan merokok, frekuensi olahraga, pola makan, konsumsi alkohol, tingkat stres, dan jam tidur. CART dipilih karena kemampuannya dalam membentuk model yang interpretatif dan menangani variabel input kategorik maupun numerik secara efektif. Pra-pemrosesan menangani nilai yang hilang, mengkodekan variabel kategorikal (jika ada), dan menskalakan fitur numerik jika diperlukan. Analisis korelasi dan bangun model *Decision Tree* (CART) untuk klasifikasi atau regresi berdasarkan variabel target yang dipilih dari dataset. Dengan menentukan variable yang akan menjadi target (dependen) dan variable yang akan digunakan sebagai fitur *(independen)* untuk membangun model. Variabel target sesuai dengan jenis tugas (kategorikal untuk klasifikasi, numerik untuk regresi).
Variabel Numerik: Dataset ini mencakup variabel numerik seperti Usia, Tinggi (cm), Berat (kg), BMI, Tingkat Stres, dan Jam Tidur. Analisis univariat menunjukkan distribusi yang bervariasi untuk setiap variabel, dengan beberapa variabel menunjukkan distribusi yang mendekati normal sementara yang lain mungkin miring atau memiliki nilai ekstrem (outlier) seperti yang terlihat pada box plot.
Variabel Kategorikal: Variabel kategorikal meliputi Jenis Kelamin, Perokok, Frekuensi Olahraga, Kualitas Diet, Konsumsi Alkohol, dan Penyakit Kronis. Distribusi frekuensi untuk variabel-variabel ini bervariasi, menunjukkan proporsi yang berbeda dalam setiap kategori.
Korelasi data set:

1. Terdapat korelasi positif yang kuat antara Berat (kg) dan BMI (koefisien korelasi Pearson sekitar 0.8591).
2. Ada korelasi negatif moderat antara Tinggi (cm) dan BMI (koefisien korelasi Pearson sekitar -0.5005).

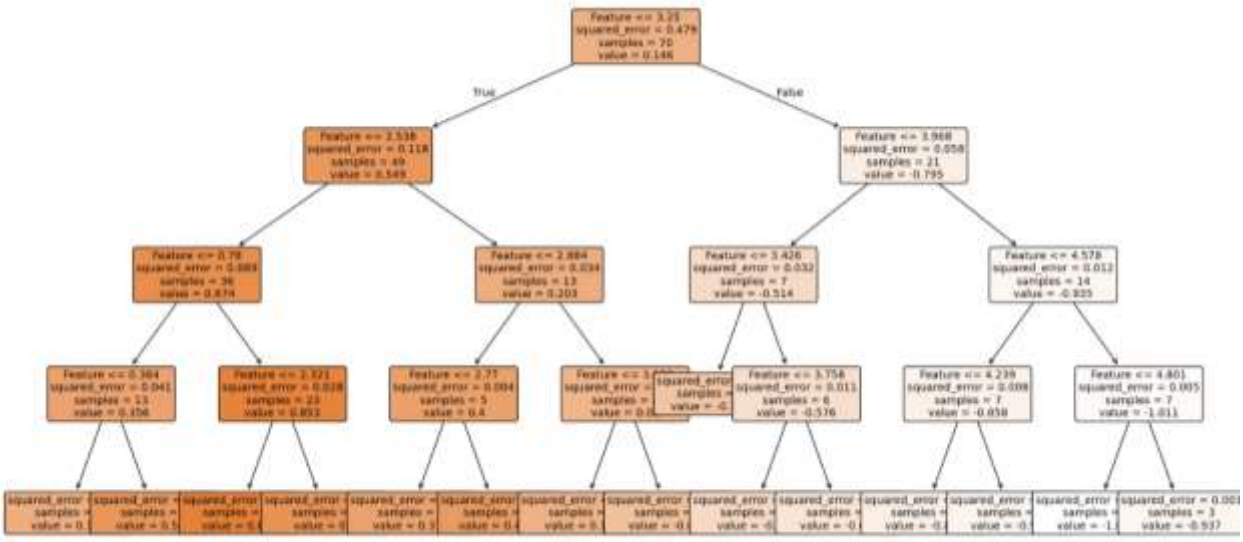
JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi

3. Tidak ada korelasi yang signifikan antara Usia dan Tingkat Stres (koefisien korelasi sekitar 0.0139).
4. Tidak ada korelasi yang signifikan antara Jam Tidur dan Tingkat Stres (koefisien korelasi sekitar 0.0015).

Node ini memisahkan data menjadi dua berdasarkan apakah nilai dari fitur tersebut lebih kecil atau sama dengan 3.25 (True) atau lebih besar (False). Masing-masing cabang kemudian terus dibagi lagi secara rekursif berdasarkan fitur dan ambang batas lainnya. Kiri (True): data dengan Feature <= 3.25 (49 sampel, rata-rata nilai = 0.549). Kanan (False): data dengan Feature > 3.25 (21 sampel, rata-rata nilai = -0.795).

Node Akar (Root Node) :

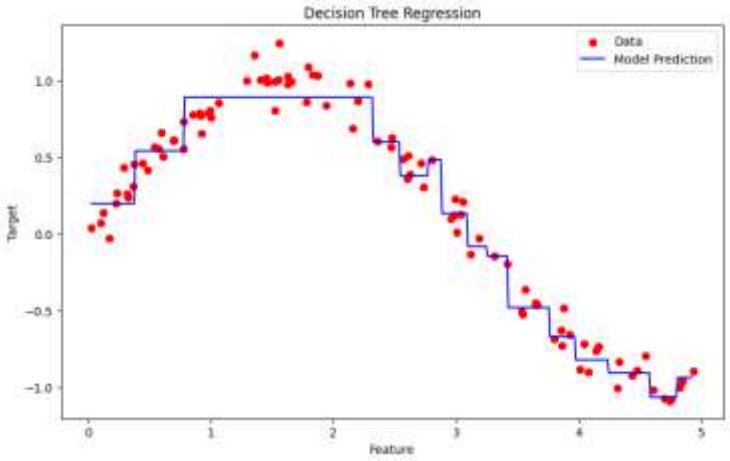
1. Baris paling atas: Feature <= 3.25
2. Squared_error = 0.479: galat kuadrat rata-rata di node ini (indicator ketidakhomogenan).
3. Samples = 70: jumlah sampel pada node ini.
4. Value = 0.146: rata-rata nilai target dari sampel-sampel.



Gambar 4. Struktur *Tree* Klasifikasi

Diperoleh hubungan yang kuat antara Berat dan BMI serta hubungan negatif moderat antara Tinggi dan BMI sesuai dengan formula perhitungan BMI, menunjukkan konsistensi data. Struktur pohon keputusan model membagi ruang fitur, menunjukkan aturan keputusan di setiap simpul, dan pohon mempartisi data untuk membuat klasifikasi. Setiap Node menampilkan Feature split menunjukkan fitur dan batas nilai yang digunakan untuk pemisahan. Squared_error menunjukkan galat kuadrat pada node. Samples menunjukkan jumlah data yang masuk ke node. Value menunjukkan nilai prediksi di node itu, yaitu rata-rata target dari data dalam node. Leaf node adalah node terminal yang tidak dibagi lag hanya berisi informasi squared_error, samples, dan value. Semua data dengan nilai fitur ≤ threshold masuk ke cabang kiri, sisanya ke kanan. Semakin kecil nilai squared_error, semakin homogen nilai target. leaf (daun), biasanya sangat kecil bahkan nol jika hanya 1 data. Jumlah total data (sampel) yang masuk ke node ini. Ini penting untuk menilai kekuatan atau kepercayaan terhadap rata-rata pada node. Hanya ada 1 sampel, dengan nilai target = 0.3, sehingga tidak ada variansi (squared_error = 0). Jika data baru masuk ke cabang ini, model akan memprediksi 0.3 sebagai nilai regresinya. Pemilihan fitur dan threshold pada tiap node dilakukan dengan minimizing squared error (MSE) antara node induk dan dua anaknya. Model seperti ini tidak menangani variabel target kategorikal. Decision Tree Regressor membagi rentang fitur menjadi segmen-segmen dan memberi prediksi konstan dalam tiap segmen. Decision Tree Regression tidak menghasilkan garis halus seperti regresi linier atau polinomial, tetapi garis tangga (piecewise constant function).

JURNAL SISTEM INFORMASI TGD
Volume 4, Nomor 4, Juli 2025, Hal 1028-1036
P-ISSN : 2828-1004 ; E-ISSN : 2828-2566
https://ojs.trigunadharma.ac.id/index.php/jsi



Gambar 5. Model Prediction Decision Tree Regression

Sumbu X – Feature menampilkan nilai variabel independen (fitur). Sumbu Y mengarah pada Target. Nilai dari variabel dependen atau target yang ingin diprediksi oleh model regresi. Setiap tangga mewakili leaf node dalam struktur pohon. Prediksi pada setiap interval adalah rata-rata target dari data dalam interval. Model mengikuti pola data dengan cukup baik. Area antara X ≈ 0.5 sampai X ≈ 2.5, model berhasil menangkap bentuk parabola dari target. Setelah X > 2.5, model tampak terlalu mengikuti noise pada data, ditunjukkan oleh banyaknya segmen kecil pada garis biru, model terlalu kompleks dan belajar pola dari data yang seharusnya dianggap noise.

4. KESIMPULAN

Penerapan algoritma *Classification and Regression Tree* (CART) untuk mengklasifikasikan gaya hidup sehat dengan pendekatan *User-Based Classification*. Data yang digunakan mencakup atribut pengguna seperti usia, jenis kelamin, BMI, kebiasaan merokok, olahraga, pola makan, konsumsi alkohol, stres, dan tidur. CART digunakan untuk membentuk pohon keputusan yang membagi data ke dalam kelas gaya hidup secara optimal. Hasil menunjukkan bahwa model mampu mengidentifikasi faktor-faktor penting yang memengaruhi gaya hidup sehat dan memberikan akurasi klasifikasi yang baik. *Decision Tree* untuk regresi, yaitu model yang memprediksi nilai numerik (target) berdasarkan nilai fitur masukan (feature), bukan klasifikasi kelas. Setiap node dalam *Decision Tree Regresi* menyajikan informasi Feature <= threshold, squared_error, samples dan value. *Decision Tree Regression* tidak menghasilkan garis halus seperti Regresi Linier atau polinomial, tetapi garis tangga *(piecewise constant function)*.

DAFTAR PUSTAKA

[1] M. I. Syafir and M. S. Saleh, "Metode partisipatif dalam meningkatkan taraf kesehatan masyarakat," vol. 02, no. 01, pp. 8–14, 2025.
[2] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024, doi: 10.30997/karimahtauhid.v3i2.11952.
[3] S. Aldana and J. S. Wibowo, "Penerapan Data Mining Terhadap Klasifikasi Pasien Penderita Penyakit Liver Menggunakan Metode K-Nearest Neighbor," *Progresif J. Ilm. Komput.*, vol. 20, no. 1, p. 124, 2024, doi: 10.35889/progresif.v20i1.1376.
[4] B. L. Fauzan, T. Agustin, and A. M. H. Mahmudah, "Prediksi Klasifikasi Kecelakaan Lalu Lintas di Kota Surakarta dengan Menggunakan Metode Regresi Logistik Multinomial," *Sustain. Civ. Build. Manag. Eng.*, vol. 1, no. 4, p. 9, 2024, doi: 10.47134/scbmej.v1i4.3159.
[5] P. Kinerja, M. Klasifikasi, and C. Saliva, "Perbandingan kinerja metode klasifikasi citra saliva ferning untuk deteksi masa subur berbasis machine learning," vol. 6, no. 2, pp. 1–8, 2023.
[6] A. Alcacer, I. Epifanio, J. Valero, and A. Ballester, "Combining classification and user-based collaborative filtering for matching footwear size," *Mathematics*, vol. 9, no. 7, pp. 1–15, 2021, doi: 10.3390/math9070771.
[7] E. Pranadjaya, E. S. Pangestu, C. O. Sereati, S. Octaviani, and M. Darmawan, "Perbandingan Algoritma Machine Learning menggunakan Orange Data Mining untuk Klasifikasi Jenis Kendaraan pada Sistem Tilang Digital," *J. Elektro*, vol. 17, no. 1, pp. 41–47, 2024, doi: 10.25170/jurnalelektro.v17i1.5429.
[8] N. Aini, M. Arif, I. T. Agustin, and Z. B. Toyibah, "Implementasi Algoritma Random Forest untuk Klasifikasi Bidang MSIB di Prodi Pendidikan Informatika," *J. Inform.*, vol. 11, no. 1, pp. 11–16, 2024, doi: 10.31294/inf.v11i1.20637.
[9] R. D. Adiningtiyas, A. Salma, and F. Fitri, "Implementation of CART Method with SMOTE for Household Poverty

Classification in Mentawai Islands 2023," vol. 2, no. 2010, pp. 422–429, 2024.

[10] N. U. Khan, W. Wan, R. Riaz, S. Jiang, and X. Wang, "Prediction and Classification of User Activities Using Machine Learning Models from Location-Based Social Network Data," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13063517.

[11] N. Ayuningtyas and W. Yustanti, "Semi-Supervised Learning pada Pelabelan dalam Klasifikasi Multi-Label Data Teks," *J. Informatics Comput. Sci.*, vol. 06, no. 1, pp. 240–248, 2024.

[12] A. Suryana, A. Irma Purnamasari, and I. Ali, "Mengoptimalkan Kepuasan Pengguna: Analisis Sentimen Review Aplikasi Grab Di Indonesia," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3396–3404, 2024, doi: 10.36040/jati.v8i3.9688.

[13] N. Pratiwi and Y. Setyawan, "Analisis Akurasi Dari Perbedaan Fungsi Kernel Dan Cost Pada Support Vector Machine Studi Kasus Klasifikasi Curah Hujan Di Jakarta," *J. Fundam. Math. Appl.*, vol. 4, no. 2, pp. 203–212, 2021, doi: 10.14710/jfma.v4i2.11691.

[14] I. Nabila, I. M. Sumertajaya, and M. Raharjo, "Penerapan Metode CART pada Pengklasifikasian Bekerja dan Pengangguran di Kabupaten Subang," *Xplore J. Stat.*, vol. 11, no. 2, pp. 120–129, 2022, doi: 10.29244/xplore.v11i2.890.

[15] M. I. Siahaan and E. Rosmaini, "Use of Classification and Regression Tree (CART) Method for Classification of Labor Force Participation Levels in Medan City in 2019," *FARABI J. Mat. dan Pendidik. Mat.*, vol. 5, no. 2, pp. 95–103, 2022, doi: 10.47662/farabi.v5i2.386.

[16] I. Nawawi and Z. Fatah, "Penerapan Decision Trees dalam Mendeteksi Pola Tidur Sehat Berdasarkan Kebiasaan Gaya Hidup," vol. 2, no. 4, pp. 34–41, 2024.

[17] W. S. MA, S. Dur, and R. Aprilia, "Analisis Waktu Kelulusan Mahasiswa Menggunakan Bagging Cart pada Fakultas Sains dan Teknologi UIN Sumatera Utara," *FARABI J. Mat. dan Pendidik. Mat.*, vol. 6, no. 2, pp. 172–179, 2023, doi: 10.47662/farabi.v6i2.628.

[18] E. Setiawaty, F. M. Afendi, and C. Suhaeni, "Metode Cart Untuk Mengidentifikasi Faktor-Faktor Yang Memengaruhi Waktu Pembelian Kendaraan Kedua," *Xplore J. Stat.*, vol. 10, no. 2, pp. 140–151, 2021, doi: 10.29244/xplore.v10i2.237.