Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 4 (2026) pp: 9616-9623
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Prediksi Status Kesehatan Berdasarkan Gaya Hidup Menggunakan Metode Decision Tree dan Feature Importance

Fajar Ramadhan[1], Dio Herlambang[2], Ambrusius Paska Dipta[3]
[1,2]Teknologi Informasi, Teknik & Informatika, Universitas Bina Sarana Informatika
fajarram055@gmail.com

Abstrak

Dalam kehidupan ini, gaya hidup dapat menjadi faktor penentu kualitas hidup seseorang, baik secara jasmani maupun secra rohani. Beberapa faktor yang dapat secara signifikan mempengaruhinya diantaranya seperti nutrisi, kebiasaan merokok, aktivitas fisik, tingkat stres, dan kesadaran diri(mindfullness), merupakan indikator penting yang mencerminkan bagaimana seseorang menjalani kesehariannya . Seiring dengan perkembangan teknologi machine learning yang menjadi sangat pesat di era saat ini,faktor-faktor tersebut dapat di evaluasi secara cepat dan akurat. Studi ini bertujuan untuk memprediksi status kesehatan berdasarkan beberapa indikator gaya hidup menggunakan metode Decision tree. Selain menghasilkan prediksi, model ini juga digunakan untuk mengevaluasi kontribusi relatif setiap fitur melalui analisis Feature importance, sehingga memperoleh pemahaman tentang faktor gaya hidup mana yang memiliki pengaruh terbesar terhadap status kesehatan. Model Decision Tree dianalisis menggunakan scikit-learn dengan pendekatan klasifikasi multikelas (Good,Average,Poor). akurasi yang dihasilkan oleh model dalam penelitian ini memperoleh hasil 80,68% dengan nilai makro F1 sebesar 0,72, yang menunjukkan kinerja stabil di semua kelas. Analisis kepentingan fitur menunjukkan bahwa kesadaran diri (mindfulness) adalah faktor yang paling dominan, diikuti oleh faktor nutrisi dan kebiasaan merokok, sementara fitur gaya hidup lainnya masih berkontribusi pada prediksi keseluruhan. Hasil ini menunjukkan bahwa decision tree mampu menjadi metode yang efisien,interaktif, dan tidak sulit untuk di interpretasikan dalam penilaian kesehatan berbasis data. Lebih lanjut, temuan mengenai faktor dominan dapat menjadi dasar untuk memberikan rekomendasi perubahan gaya hidup guna meningkatkan kesehatan individu.

***Kata kunci**: Decision Tree, Prediksi Kesehatan, Gaya Hidup, Feature Importance, Machine Learning*

1.Latar Belakang

Kesehatan merupakan aspekyang sangat penting dalam kehidupan seseorang dan membutuhkan perhatian yang cermat, karena dapat dipengaruhi oleh berbagai faktor, terutama faktor gaya hidup. nutrisi, kebiasaan merokok,aktivitas fisik, stres, dan kesadaran diri adalah indikator bagaimana seseorang mengelola gaya hidupnya. Dengan perkembangan teknologi *machine learning*, analisis faktor-faktor perilaku ini dapat dilakukan dengan lebih akurat dan efisien untuk menilai status kesehatan seseorang. Pendekatan ini memungkinkan proses evaluasi kesehatan yang lebih objektif dan berbasis data, mendukung pengambilan keputusan dan rekomendasi untuk perbaikan gaya hidup.

Penelitian yang telah ada sebelumnya yang berjudul Penerapan Decision Tree untuk Klasifikasi Status Kesehatan dengan perbandingan KNN dan Naive Bayes sudah menerapkan metode *machine learning* untuk memprediksi kondisi kesehatan, salah satunya menggunakan dataset yang masih berupa dataset dummy yang dibuat secara sederhana dengan Ai yang mana dengan jumlah data yang terbatas. Penelitian tersebut menggunakan beberapa metode, termasuk Decision Tree, namun fitur yang digunakan masih sangat minim, evaluasi modelnya juga masih terbilang sederhana, dan dataset dummy tidak merepresentasikan kondisi nyata.(Biyantoro & Prasetiyo, 2024).hal ini yang menyebabkan, akurasi yang dihasilkan sangatlah tinggi tetapi tidak mencerminkan performa model pada data yang lebih jelas & kompleks.sebagai perbandingannya , penelitian ini menggunakan dataset nyata yang bersumber dari kaggle dengan jumlah data jauh lebih besar dan indikator gaya hidup yang lebih besar, sehingga menghasilkan analisis yang lebih valid dan realistis. Pembeda antara penelitian terdahulu dan penelitian ini:

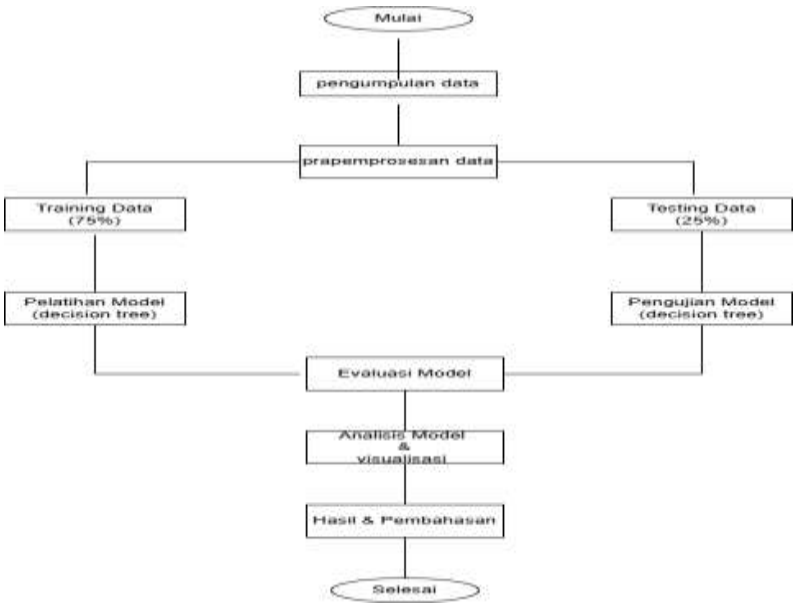
Tabel 1 Perbedaan Penelitian terdahulu dan Penelitian Ini

Aspek Pembanding	Penelitian terdahulu	Penelitian ini
Jenis dan jumlah Dataset	Dummy,buatan AI,tidak realistis Sangat sedikit (120 baris)	Dataset berasal dari kaggle, besar, representatif (10.000 baris)
Jumlah fitur	4 fitur sederhana (Age, Gender, Weight, ExerciseHours)	Banyak fitur gaya hidup (Mindfulness, Nutrition, Smoking, dll.)
Metode yang digunakan	DT, KNN, Naive Bayes (tanpa analisis mendalam)	Decision Tree (analisis mendalam + feature importance)
Evaluasi Model	Hanya accuracy dan Confusion Matrix sederhana	Accuracy, Precision, Recall, Macro F1-score, Confusion Matrix lengkap
Interpretasi Model	Hanya menampilkan aturan pohon (rule-based)	Feature Importance, penjelasan faktor dominan, interpretasi pohon
Kualitas Data	Data dummy → pola mudah ditebak → akurasi tinggi tidak realistis	Data nyata → pola kompleks → akurasi lebih wajar dan valid
Tujuan Analisis	Sekadar membandingkan beberapa algoritma dasar	Fokus interpretasi faktor gaya hidup yang mempengaruhi kesehatan

Dari tabel yang digunakan untuk perbedaan tersebut menunjukan bahwa penelitian terdahulu memiliki beberapa keterbatasan, terutama terkait ukuran dataset, kualitas data, jumlah fitur, serta kedalaman evaluasi model. Hal ini lah yang dapat menyebabkan hasil prediksi terlihat baik pada data sederhana tetapi tidak mencerminkan performa model pada data nyata. Penelitian ini dilakukan untuk mengatasi keterbatasan tersebut dengan memanfaatkan dataset nyata berukuran besar, indikator gaya hidup yang lebih komprehensif, serta evaluasi model yang lebih lengkap dan interpretatif melalui pengunaan analisis *feature importance*. Penelitian ini difokuskan pada pengembangan sebuah model klasifikasi dengan pengunaan metode Decision Tree untuk memprediksi status kesehatan suatu individu berdasarkan indikator gaya hidup, sekaligus untuk memahami kontribusi masing-masing fitur terhadap prediksi yang dihasilkan. Penelitian ini diharapkan dapat memberikan gambaran yang lebih akurat dan interpretatif, sehingga dapat menjadi dasar dalam memberikan rekomendasi peningkatan kualitas kesehatan melalui perubahan gaya hidup.

2.Metode Penelitian

Dalam penelitian prediksi status kesehatan ini, di proses melewati beberapa tahapan utama yang meliputi prapemrosesan data, pemisahan dataset menjadi data yang akan di latih dan data yang akan uji, pelatihan pada model, pengujian model, serta evaluasi dan analisis hasil. Tahapan tersebut disusun secara sistematis guna memastikan bahwa model Decision Tree yang dibangun bisa memberikan prediksi yang akurat dan mudah diinterpretasikan. Berikut ini merupakan alur metode penelitian dalam bentuk flowchart



Gambar 1 proses penelitian

DOI: https://doi.org/10.31004/riggs.v4i4.5246

2.1. Pengumpulan Data

dataset kesehatan yang digunakan dalam penelitian ini berasal dari Kaggle dengan judul "Holistic Health & Lifestyle Score Dataset", yang berisi berbagai indikator gaya hidup yang relevan untuk analisis status kesehatan. Datasetterdiri dari sembilan variabel gaya hidup,dan juga termasuk tiga fitur utama yang menjadi fokus penelitian ini, yaitu *mindfulness*, *nutrition score*, dan *smoking habit*, serta satu variabel target berupa Health Status dengan tiga kategori kelas (Good, Average, dan Poor). Seluruh fitur gaya hidup yang ada dalam dataset telah berformat numerik sehingga dapat digunakan secara langsung dalam proses analisis dan pelatihan model. untuk variabel target *Health Status* direpresentasikan ke bentuk numerik melalui proses *label encoding* agar dapat diproses oleh algoritma *Decision Tree*.

2.2 Prapemrosesan Data

Tahap prapemrosesan data berguna untuk memastikan kualitas data sebelum proses pelatihan model. Pada tahap ini dilakukan pemeriksaan kelengkapan data, konsistensi format data , serta relevansi fitur yang digunakan agar tidak terjadi bias atau kesalahan dalam proses klasifikasi. Prapemrosesan yang tepat diperlukan untuk meningkatkan keandalan dan akurasi model Decision Tree yang dibangun.Beberapa tahapan yang dilakukan pada tahap ini ialah sebagai berikut:

2.2.1 Pemeriksaan Data

Dataset diperiksa untuk memastikan tidak terdapat *missing values*, data ganda, maupun nilai yang tidak jelas pada sembilan fitur utama. Seluruh variabel telah berformat numerik sehingga tidak memerlukan proses *encoding* tambahan.Kolom Overall_Health_Score yang terdapat pada dataset dihapus karena merupakan skor gabungan yang dihitung dari beberapa fitur gaya hidup. Penggunaan kolom ini sebagai input dapat menyebabkan *data leakage* dan menghasilkan prediksi yang tidak valid. Oleh karena itu, hanya fitur-fitur asli yang digunakan dalam proses pelatihan model.

2.1.2 Pemisahan Data (Train–Test Split)

Datasetnya kemudian di pisah menjadi dua bagian menggunakan fungsi train_test_split:

$$Train = 75\%, Test = 25\%$$

Pemisahan ini dilakukan menggunakan parameter stratify berdasarkan label *Health Status* untuk menjaga proporsi kelas tetap seimbang pada data yang akan dilatih dan data yang akan di uji. Hal ini bertujuan mendapatkan evaluasi performa model yang objektif.

2.2 Algoritma Decision Tree

Algoritma Decision Tree ialah metode pengklasifikasian yang bekerja memisahkan data ke dalam kelompok-kelompok yang semakin homogen melalui struktur yang berbentuk pohon. Algoritma ini dipilih karena memiliki interpretabilitas yang tinggi, proses pengambilan keputusan yang transparan, serta mampu menangani data numerik tanpa memerlukan normalisasi.Pada setiap tahap pemisahan (split), Decision Tree juga memilih fitur dan nilai ambang (threshold) yang sangat efektif dalam memisahkan data berdasarkan kelasnya. Proses pemilihan split tersebut dilakukan dengan menghitung penurunan impuritas pada setiap node, sehingga node cabang yang terbentuk semakin homogen.

.2.2.1 Indeks Gini

Impuritas node dihitung menggunakan rumus Gini:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

dengan:

a. p_i= proporsi sampel pada kelas ke–i

b. nilai Gini rendah → node lebih homogen

c. nilai Gini tinggi → data pada node lebih bercampur

DOI: https://doi.org/10.31004/riggs.v4i4.5246

2.2.2 Information Gain

Pemilihan split terbaik dilakukan berdasarkan *Information Gain (IG)*, yaitu pengurangan impuritas:

$$IG = Gini(parent) - (\frac{N_{left}}{N} \cdot Gini(left) + \frac{N_{right}}{N} \cdot Gini(right))$$

Split dengan nilai IG terbesar dipilih menjadi pemecah node.

2.2.3 Pembentukan Struktur decision tree

Proses pembentukan pohon dilakukan secara rekursif hingga memenuhi salah satu kondisi berikut:

1. impuritas tidak berkurang secara signifikan,
2. jumlah sampel pada node terlalu sedikit, atau
3. pohon mencapai batas kedalaman tertentu (*max depth*).

2.3 Feature Importance

Feature importance adalah fitur yang digunakan untuk mengukur kontribusi setiap fitur dalam proses klasifikasi. Pada Decision Tree, nilai importance dihitung berdasarkan total penurunan impuritas yang dihasilkan oleh fitur tersebut.

Rumus dasar feature importance:

$$FI_t = \frac{n_t}{N} \cdot \Delta Gini_t$$

dengan:

a. n_t= jumlah sampel pada node t
b. N= total sampel
c. $\Delta Gini_t$= penurunan impuritas setelah split

Nilai importance dari seluruh node dijumlahkan dan dinormalisasi sehingga:

$$\sum FI = 1$$

Hasil analisis digunakan untuk mengidentifikasi fitur yang paling berpengaruh terhadap prediksi status kesehatan. Dalam penelitian ini, *mindfulness* merupakan fitur dengan kontribusi terbesar.

2.4 Evaluasi pada Model

Model dievaluasi dengan penggunaan beberapa metrik standar klasifikasi:

1. Akurasi

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

Mengukur proporsi prediksi yang benar dari seluruh data uji.

DOI: https://doi.org/10.31004/riggs.v4i4.5246

2. Precision

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

Mengukur ketepatan model dalam memprediksi kelas tertentu.

3. Recall

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

Mengukur kemampuan model dalam mendeteksi data yang benar dari suatu kelas.

4. F1-score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Mengukur keseimbangan antara precision dan recall.

Alat Evaluasi yang Digunakan

Evaluasi dilakukan menggunakan:

a. Classification Report → berisi precision, recall, dan F1-score untuk tiap kelas

b. Confusion Matrix → memetakan kesalahan prediksi antar kelas

c. Macro F1-score → menghitung performa rata-rata pada kelas *Good*, *Average*, dan *Poor*

Seluruh evaluasi dilakukan pada data yang diuji untuk mengukur kemampuan generalisasi model.

3. Hasil dan Pembahasan

Tahap ini menyajikan hasil pengujian model klasifikasi menggunakan algoritma Decision Tree dalam memprediksi status kesehatan berdasarkan indikator gaya hidup.Sebanyak 25% dataset digunakan sebagai data yang diuji untuk mengevaluasi kinerja model berdasarkan indikator gaya hidup.

Hasil evaluasi performa model menandakan bahwa Decision Tree menghasilkan nilai akurasi mencapai 80,68%, yang menandakan bahwa sebagian besar data uji berhasil dalam proses pengklasifikkasian. Selain akurasi, evaluasi juga dilakukan mellaui pengunaan metrik precision, recall, dan F1-score untuk menghasilkan gambaran performa pada model yang lebih komprehensif pada setiap kelas. Nilai macro F1-score memperoleh nilai 0,72 mengindikasikan bahwa performa model relatif seimbang ketika setiap kelas diperlakukan secara setara, meskipun dataset memiliki distribusi kelas yang tidak merata. Untuk nilai weighted average F1-score sendiri telah mencapai 0,81 yang mengindikasikan bahwa model secara keseluruhan memiliki kinerja yang baik dengan mempertimbangkan proporsi jumlah data pada setiap kelas.

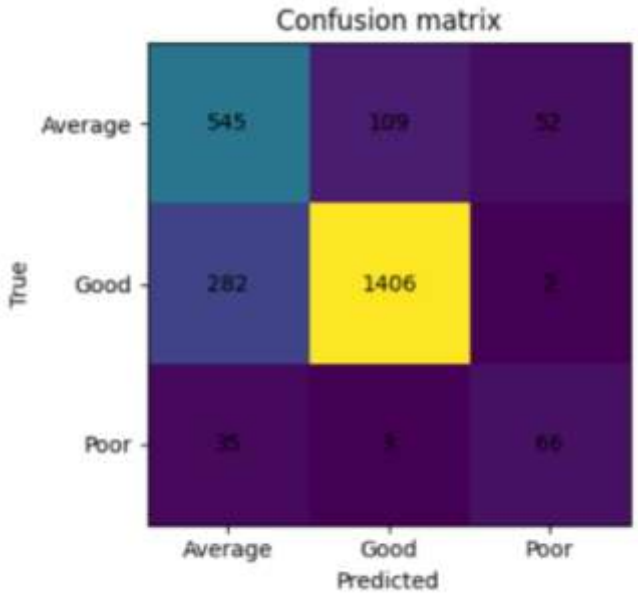
Tabel 2. Tabel Evaluasi

kelas	Precision	Recall	F1-Score	Support
Good	**0.93**	**0.84**	0.88	**887**
Average	**0.78**	**0.94**	0.86	**924**
Poor	**0.67**	**0.22**	0.35	**213**
Accuracy			80.68%	**2024**
Macro avg	**0.70**	**0.75**	0.72	**2024**
Weighted avg	**0.83**	**0.81**	0.81	**2024**

Berdasarkan hasil confusion matrix, model menunjukkan kemampuan yang baik dalam mengenali kelas Good dan Average, yang tercermin dari tingginya nilai recall kedua kelas tersebut. Hal ini mengindikasikan model mampu mengenali sebagian besar data aktual dari kelas mayoritas dengan tepat. Namun, performa model pada kelas Poor masih sangatlah terbatas terbatas,

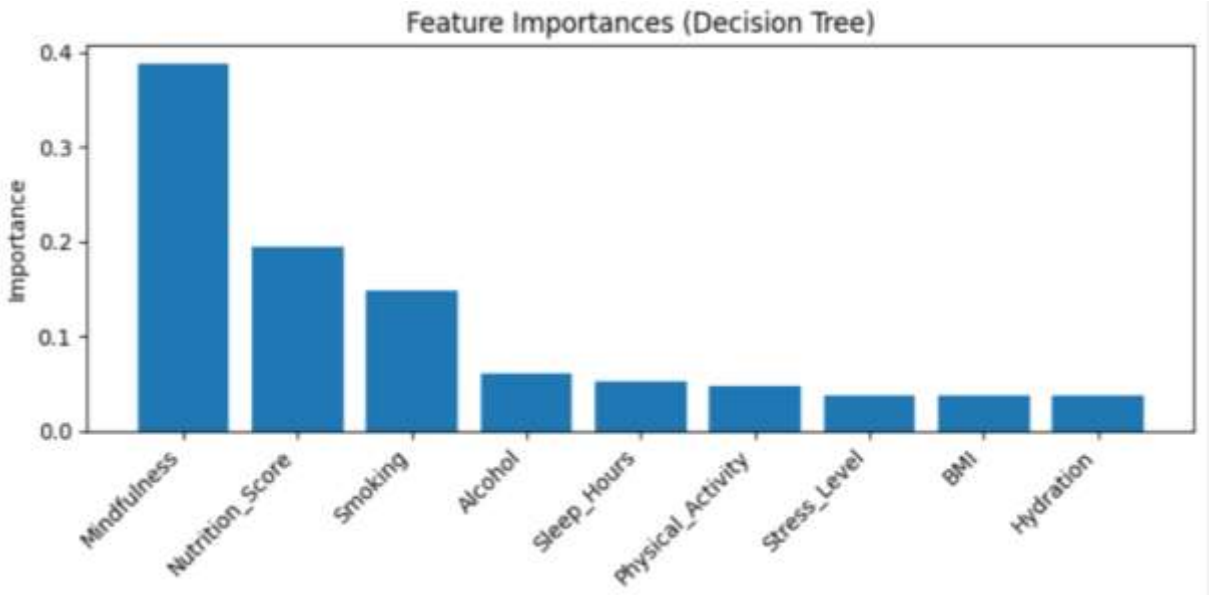
DOI: https://doi.org/10.31004/riggs.v4i4.5246

dengan nilai recall yang hanya mencapai 0,22. Rendahnya nilai ini menunjukkan bahwa model menghadapi kesulitan dalam mengenali kelas minoritas, yang dikarenakan oleh ketidakseimbangan distribusi data serta dipengaruhi juga adanya kemiripan karakteristik antara kelas Poor dan Average pada beberapa indikator gaya hidup.



Gambar 2 Hasil Evaluasi Model Confusion Matrix

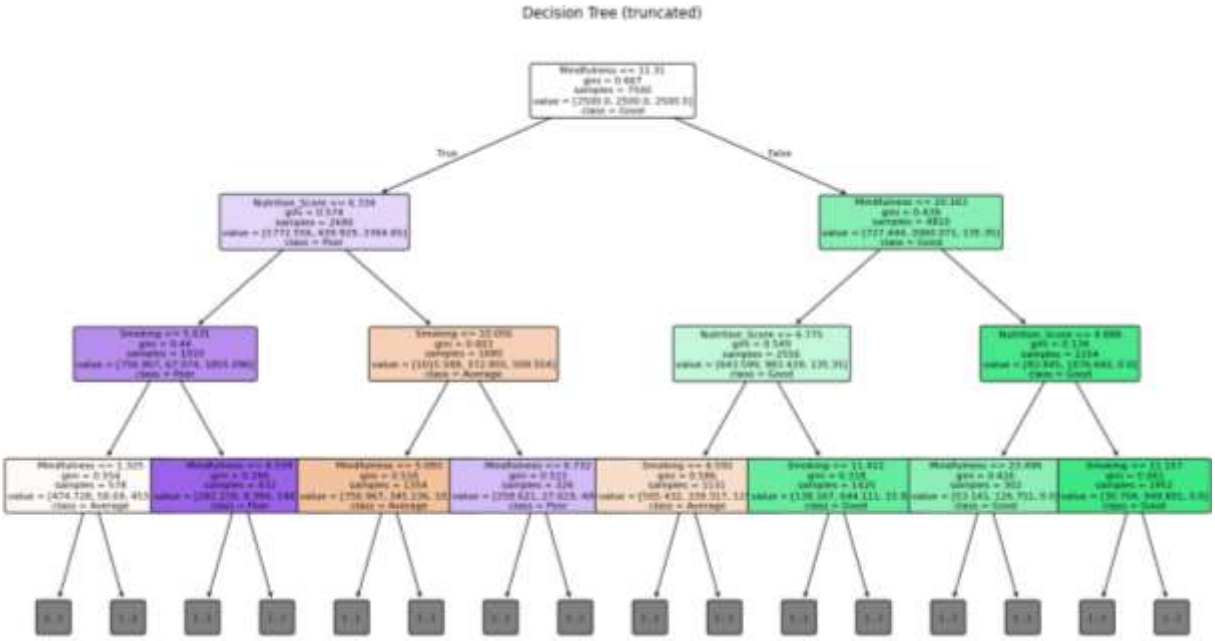
Analisis feature importance menunjukkan bahwa mindfulness merupakan fitur dengan kontribusi yang sangat mendominasi dalam proses klasifikasi status kesehatan. Temuan ini mengindikasikan bahwa faktor psikologis, khususnya kemampuan individu dalam mengelola stres dan menjaga kesadaran diri, memiliki peran penting dalam membedakan kondisi kesehatan. Fitur nutrition score berada pada urutan berikutnya, yang menunjukkan bahwa pola makan juga menjadi faktor signifikan dalam menentukan status kesehatan individu, terutama dalam membedakan kelas Average dan Poor. Sementara itu, smoking habit memiliki nilai kontribusi yang lebih rendah dibandingkan dua fitur utama tersebut, namun tetap berperan dalam mendukung keputusan model. Meskipun terdapat beberapa fitur lain dengan nilai importance yang lebih kecil, ketiga indikator utama ini secara kolektif menjadi dasar utama pengambilan keputusan model.



Gambar 3 Hasil analisis berdasarkan feature importance

DOI: https://doi.org/10.31004/riggs.v4i4.5246

Visualisasi struktur decision tree menunjukkan bahwa mindfulness digunakan sebagai node awal, yang mengindikasikan bahwa fitur ini memberikan penurunan impuritas terbesar pada saat proses pemisahan data. Cabang-cabang selanjutnya memperlihatkan peran nutrition score dan smoking habit dalam membedakan kelas Average dan Poor. Pola keputusan yang terbentuk menunjukkan bahwa individu dengan nilai mindfulness yang tinggi cenderung diklasifikasikan ke dalam kelas Good, yang menegaskan keterkaitan antara kondisi psikologis, kebiasaan hidup sehat, dan status kesehatan secara keseluruhan.



Gambar 4 hasil Dari decision Tree

hasil penelitian ini mengindikasikan algoritma Decision Tree mampu digunakan sebagai pendekatan yang efektif dan mudah diinterpretasikan dalam memprediksi status kesehatan berbasis indikator gaya hidup. Keunggulan utama model terletak pada kemampuannya dalam memberikan interpretasi yang jelas melalui analisis feature importance dan struktur pohon keputusan. Namun, penelitian ini masih memiliki keterbatasan, terutama pada ketidakseimbangan distribusi data kelas Poor yang memengaruhi kinerja model dalam mengenali kelas tersebut. Penelitian selanjutnya dapat mempertimbangkan penggunaan teknik penyeimbangan data atau pendekatan ensemble untuk meningkatkan performa klasifikasi, khususnya pada kelas minoritas.

4.Kesimpulan

Penelitian ini berhasil mengembangkan modeluntuk mengklasifikasi status kesehatan melalui penggunaan algoritma Decision Tree berdasarkan indikator gaya hidup, yaitu mindfulness, nutrition score, dan smoking habit. Model juga menghasilkan akurasi mencapai 80,68% dan nilai macro F1-score sebanyak 0,72, yang mengindikasikan performa klasifikasi berada pada tingkat yang baik meskipun dataset memiliki ketidakseimbangan kelas. Analisis feature importance menunjukkan bahwa mindfulness merupakan faktor paling berpengaruh dalam proses klasifikasi, diikuti oleh nutrition score dan smoking habit, yang menegaskan bahwa aspek psikologis, pola makan, dan perilaku hidup memiliki kontribusi signifikan dalam penentuan status kesehatan. Secara keseluruhan, model yang dikembangkan dapat menjadi alat bantu yang informatif untuk memahami hubungan antara gaya hidup dan kesehatan, serta berpotensi digunakan dalam analisis risiko kesehatan berbasis data. Penelitian selanjutnya dapat mempertimbangkan penggunaan teknik penyeimbangan data, penambahan variabel gaya hidup lainnya, atau penerapan metode ensemble seperti Random Forest untuk mengoptimalkan kinerja model, khususnya pada kelas Poor.

Referensi

1. Anissa, A. I., & Qoiriah, A. (2025). *Prediksi Tingkat Stres Berdasarkan Pola Hidup Menggunakan Machine Learning*. *07*, 292–300.
2. Artiyasa, M., & Yuda, G. S. (2025). *Penerapan H2O AutoML untuk Prediksi Kanker Kolorektal*. *4*(1), 222–232.
3. Biyantoro, A. S., & Prasetiyo, B. (2024). *Application of Decision Tree for Health Status Classification , Compared to KNN and Naive Bayes Penerapan Decision Tree untuk Klasifikasi Status Kesehatan dengan perbandingan KNN dan Naive Bayes*. *4*(1), 47–55.
4. Informatika, P. S., & Mandiri, U. N. (2022). *Metode Naïve Bayes Untuk Memprediksi Penyakit Stroke*.

DOI: https://doi.org/10.31004/riggs.v4i4.5246

5. Irwansyah, B., Jolyarni, N., Riswan, D., & Damanik, S. (2025). *Volume 3 Nomor 3 Agustus 2025 DOI : https://doi.org/10.62027/sevaka.v3i3.554 PENYULUHAN PREDIKSI RISIKO RAMBUT RONTOK MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)*
6. Karuniasari, W., & Prathivi, R. (2025). *Komparasi Algoritma Random Forest dan XGBoost dalam Prediksi Premi Asuransi Kesehatan*. *23*(1), 118–125.
7. Meylani, A., Negara, E. S., Darma, U. B., & Selatan, S. (2022). *Aplikasi Prediksi Kesehatan Menggunakan*. *14*(2), 208–215.
8. Miah, A. (2025). *Holistic Health & Lifestyle Score Dataset*. Kaggle. https://www.kaggle.com/datasets/miadul/holistic-health-and-lifestyle-score-dataset
9. Nawawi, I., & Fatah, Z. (2024). *Penerapan Decision Trees dalam Mendeteksi Pola Tidur Sehat Berdasarkan Kebiasaan Gaya Hidup*. *2*(4), 34–41.
10. Novari, A. S., & S, U. K. N. (2024). *Prediksi Faktor yang Mempengaruhi Hipertensi dengan Metode Data Mining untuk meningkatkan Pelayanan Kesehatan di UPT Puskesmas Ngoro*. 1–16.
11. Prakoso, R. N., Rochim, S. I., Subarna, A., & K, M. E. (2025). *Perbandingan Algoritma Naïve Bayes Dan Random Forest Dalam Klasifikasi Obesitas Berdasarkan Faktor Gaya Hidup*. *09*, 11–18.
12. Prianto, C., Angelina, R., & Hutabarat, P. (2025). *Penerapan Algoritma Machine learning untuk Prediksi Gangguan Kesehatan Mental : Systematic Literature Review Application of Machine learning Algorithms for Predicting Mental Health Disorders : A Systematic Literature Review*. *13*(4), 510–518. https://doi.org/10.26418/justin.v13i4.95911
13. Rahayu, C. A., Hartono, R., & Sudiarjo, A. (2023). *PREDIKSI PENDERITA DIABETES MENGGUNAKAN*. *11*(3).
14. Sabna, E., & Dewi, O. (2025). *Prediksi Penyakit Stroke menggunakan Algoritma Decision Tree dan Naïve Bayes*. *4*(3), 1294–1299.
15. Sitohang, C., Ginting, F. E., Br, Y. M., Masyarakat, K., & Publik, K. (2024). *Prediksi Jumlah Perokok dan Dampaknya terhadap Kesehatan Masyarakat Menggunakan Regresi Linear*. *1*(2), 512–516.
16. Wafa, H. S., Hadiana, A. I., Umbara, F. R., Terusan, J., Sudirman, J., Sel, K. C., & Cimahi, K. (2022). *Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)*. *1*, 40–45.

DOI: https://doi.org/10.31004/riggs.v4i4.5246