JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 1-6
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Perbandingan Linear Regression, K-NN, dan Naive Bayes untuk Prediksi Kategori BMI

Adam Wildan Firdaus[1], Mochamad Danu Rizkyarto[2], Rizky Rachma Putri[3]

[1,2,3]Teknologi Informasi, Universitas Bina Sarana Informatika, Depok, Indonesia
Email: [1]17230123@bsi.ac.id, [2]17230112@bsi.ac.id, [3]17230428@bsi.ac.id

Abstrak−Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja tiga algoritma Machine Learning, yaitu *Linear Regression, K-Nearest Neighbor,* dan *Gaussian Naive Bayes* dalam memprediksi kategori Indeks Massa Tubuh (*BMI*) berdasarkan data *antropometri*. Dataset yang digunakan terdiri dari 111 sampel yang diperoleh dari repositori Zenodo dan diproses melalui tahapan pra-pemrosesan, standarisasi fitur numerik, serta evaluasi menggunakan metode *Stratified K-Fold Cross Validation*. Hasil pengujian menunjukkan adanya perbedaan karakteristik pada masing-masing algoritma. *Gaussian Naive Bayes* memberikan akurasi tertinggi dalam klasifikasi kategori *BMI* sebesar 75,74%, sedangkan *Linear Regression* menghasilkan prediksi numerik paling presisi dengan nilai *Mean Absolute Error* 0,362 kg/m². *K-Nearest Neighbor* memiliki akurasi lebih rendah pada dataset kecil, namun pada uji data pengguna baru justru memberikan hasil prediksi *BMI* yang paling mendekati nilai aktual. Temuan ini menegaskan bahwa tidak ada satu algoritma yang unggul di semua aspek, melainkan setiap metode memiliki kelebihan sesuai konteks penggunaannya. Oleh karena itu, pemilihan algoritma sebaiknya mempertimbangkan tujuan aplikasi, apakah menekankan ketepatan angka *BMI* atau klasifikasi kategori secara cepat dan efisien.

Kata Kunci: Indeks Massa Tubuh, Machine Learning, Gaussian Naive Bayes, Linear Regression, K-Nearest Neighbor

Abstract− This study aims to analyze and compare the performance of three *Machine Learning* algorithms, namely *Linear Regression, K-Nearest Neighbor,* and *Gaussian Naive Bayes*, in predicting *Body Mass Index (BMI)* categories based on anthropometric data. The dataset used consists of 111 samples obtained from the *Zenodo* repository and was processed through preprocessing, feature standardization, and evaluation using the *Stratified K-Fold Cross Validation* method. The results reveal distinct characteristics for each algorithm. *Gaussian Naive Bayes* achieved the highest accuracy in *BMI* category classification at 75.74%, while *Linear Regression* produced the most precise numerical predictions with a *Mean Absolute Error* of 0.362 kg/m². *K-Nearest Neighbor* showed lower accuracy on the small dataset (66.90%), but in testing with new user data it provided the closest prediction to the actual *BMI* value. These findings highlight that no single algorithm outperforms the others in all aspects; rather, each has strengths depending on the context of use. Therefore, the choice of algorithm should be aligned with the application's objective, whether prioritizing precise numerical *BMI* estimation or rapid categorical classification.

Keywords: Body Mass Index, Machine Learning, Gaussian Naive Bayes, Linear Regression, K-Nearest Neighbor

1. PENDAHULUAN

Kasus penyakit tidak menular (*Non-Communicable Diseases/NCDs*) terus mengalami peningkatan, menjadi masalah kesehatan serius di berbagai tempat. Laporan Organisasi Kesehatan Dunia menunjukkan bahwa penyakit ini menyebabkan kematian terbanyak secara global, dengan jumlah mencapai perkiraan 36 juta orang per tahun [1]. NCDs tidak hanya menimbulkan beban biaya yang besar bagi layanan kesehatan, tetapi juga menyebabkan penurunan kualitas hidup, hilangnya kemampuan kerja, serta menimbulkan masalah sosial dan ekonomi pada individu, keluarga, dan masyarakat. Salah satu penyebab utamanya adalah kelebihan berat badan hingga obesitas, yang dapat diukur melalui patokan berat badan ideal yang disebut Indeks Massa Tubuh (*Body Mass Index/BMI*). *BMI* merupakan metode ukur sederhana yang digunakan untuk menilai kondisi kesehatan seseorang berdasarkan tinggi dan berat badannya, serta untuk mengidentifikasi risiko terkena penyakit jangka panjang seperti diabetes dan hipertensi [2].

Prevalensi obesitas di Indonesia terus meningkat hingga menjadi perhatian. Data Survei Kesehatan Indonesia [3] menjelaskan masalah kelebihan gizi pada orang dewasa cukup tinggi, yang mengindikasikan perlunya langkah penanganan yang tepat. Faktor pendorongnya antara lain pola makan tinggi kalori, kurangnya aktivitas fisik, dan rendahnya kesadaran untuk memantau berat badan dengan standar *BMI* secara rutin. Kondisi ini diperparah oleh keterbatasan fasilitas di layanan kesehatan tingkat dasar dan kurangnya akses terhadap alat diagnosis yang cepat dan mudah dijangkau. Oleh karena itu, pengembangan sistem untuk memprediksi *BMI* secara akurat dan mudah diakses menjadi hal yang penting untuk mendukung pencegahan dini dan manajemen kesehatan yang lebih baik, khususnya di layanan kesehatan primer dan platform kesehatan digital.

Oleh karena itu, penggolongan kategori *BMI* secara cepat dan akurat berdasarkan informasi dasar seperti berat badan, tinggi badan, jenis kelamin, dan usia sangat penting sebagai langkah pemeriksaan kesehatan dini [4]. Dalam konteks ini, pemanfaatan *Machine Learning* menjadi sangat relevan karena kemampuannya untuk mengotomasi proses klasifikasi dan prediksi data kesehatan secara lebih efisien dan akurat. *Machine Learning* mampu mengolah volume data yang besar dan kompleks, mengidentifikasi pola yang sulit dikenali yang mungkin terlewat oleh metode statistik tradisional, serta memberikan gambaran awal untuk membantu pengambilan keputusan yang dapat menunjang

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **30/11/2025**; Accepted: **01/01/2026**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 1-6
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

pengambilan keputusan praktik pelayanan kesehatan. Penelitian ini membandingkan kemampuan tiga metode *Machine Learning*, yaitu *Linear Regression, K-Nearest Neighbor*, dan *Gaussian Naive Bayes* dalam memprediksi kategori *BMI*.

Meskipun sejumlah penelitian telah menerapkan berbagai algoritma *Machine Learning* untuk mengatasi masalah obesitas dan klasifikasi status gizi termasuk studi yang membandingkan algoritma populer seperti *K-Nearest Neighbor* dan *Gaussian Naive Bayes* [5], serta penelitian yang menunjukkan kekuatan regresi linier untuk memodelkan nilai *BMI* secara berkelanjutan [6], masih sedikit studi komparatif yang menempatkan *Linear Regression* sebagai acuan dasar atau titik awal perbandingan lalu langsung membandingkannya dengan metode klasifikasi non-linier seperti *K-Nearest Neighbor* dan *Gaussian Naive Bayes* pada tugas klasifikasi kategori *BMI* menggunakan penilaian yang sama. Selain itu, beberapa penelitian menggunakan dataset dan langkah awal pengolahan yang berbeda sehingga sulit melakukan perbandingan kinerja yang adil antar model [7]. Penelitian lain juga telah mengembangkan model yang lebih kompleks seperti *Multilayer Perceptron* (MLP) untuk prediksi *BMI* dan variabel kesehatan lainnya [8], serta *Decision Tree* untuk klasifikasi obesitas [9]. Penelitian ini melengkapi kekurangan dalam penelitian sebelumnya dengan melakukan perbandingan terarah antara *Linear Regression, K-Nearest Neighbor,* dan *Gaussian Naive Bayes* pada dataset yang sama, dengan langkah awal pengolahan konsisten dan penilaian menyeluruh menggunakan akurasi klasifikasi serta ukuran kesalahan dalam prediksi angka seperti *Mean Absolute Error (MAE)* dan *Root Mean Square Error (RMSE)* untuk menentukan model yang paling efisien dan praktis untuk aplikasi layanan kesehatan primer.

Kondisi tersebut menunjukkan bahwa sebagian besar penelitian sebelumnya masih memisahkan pendekatan regresi dan klasifikasi dalam prediksi *BMI*, sehingga belum memberikan gambaran yang jelas mengenai perbandingan kinerja model sederhana dan model yang lebih kompleks pada konteks dan skema evaluasi yang sama. Akibatnya, pemahaman mengenai kelebihan dan keterbatasan masing-masing pendekatan dalam menentukan kategori *BMI* yang praktis untuk aplikasi kesehatan masih terbatas. Kondisi ini berpotensi menyulitkan pemilihan metode yang paling tepat untuk diterapkan pada sistem prediksi BMI yang sederhana, cepat, dan mudah diimplementasikan pada layanan kesehatan primer.

Algoritma seperti *K-Nearest Neighbor* dan *Gaussian Naive Bayes* sudah banyak digunakan dan terbukti memberikan hasil yang baik dalam klasifikasi data kesehatan. Misalnya, penelitian oleh Lemantara & Lusiani [10] yang memprediksi penyakit diabetes pada perempuan dengan memasukkan nilai *BMI* sebagai salah satu fitur, menunjukkan akurasi sebesar 77,98% untuk *K-Nearest Neighbor* dan 78,35% untuk *Gaussian Naive Bayes*. Penelitian lain juga menunjukkan bahwa algoritma yang lebih kompleks seperti *Support Vector Machine* dapat mencapai akurasi hingga sekitar 95% dalam menentukan kategori *BMI* [11]. Namun, masih sedikit penelitian yang menggunakan *Linear Regression* sebagai model dasar atau pembanding (*baseline*) dalam klasifikasi *BMI* yang memiliki beberapa kategori. Sebagian besar studi langsung menggunakan algoritma yang lebih rumit tanpa melihat sejauh mana model linear sederhana dapat memberikan hasil yang mendekati. Padahal, model dasar seperti *Linear Regression* penting digunakan sebagai acuan untuk mengetahui seberapa besar peningkatan kinerja yang sebenarnya diberikan oleh algoritma yang lebih kompleks. Perbandingan ini menjadi krusial untuk memahami pertimbangan antara tingkat kerumitan model, kemudahan pemahaman, dan hasil yang diperoleh.

Dalam penelitian ini, *Linear Regression* digunakan sebagai acuan dasar untuk menggambarkan hubungan linier sederhana antara variabel masukan atau input dan nilai *BMI*. Prediksi nilai *BMI* kontinu kemudian akan dikonversi menjadi kategori diskrit menggunakan batas nilai tertentu. Pendekatan ini dipilih karena kesederhanaan, kecepatan, dan kemudahannya dalam menginterpretasi pengaruh tiap fitur, menjadikannya titik awal yang ideal untuk perbandingan. Sementara itu, *K-Nearest Neighbor* digunakan karena kemampuannya menangkap pola non-linier berdasarkan kedekatan jarak tanpa membuat asumsi distribusi data yang ketat, yang sangat relevan untuk data kesehatan yang seringkali tidak terdistribusi secara normal [12]. Di sisi lain, *Gaussian Naive Bayes* dipilih karena efisiensinya pada dataset berukuran kecil hingga menengah [13]. Perbandingan ketiganya, dari yang paling sederhana menggunakan *Linear Regression* hingga yang telah teruji di domain serupa dengan *K-Nearest Neighbor* dan *Naive Bayes* dalam konteks yang sama, merupakan kontribusi spesifik dari penelitian ini untuk memberikan pemahaman yang lebih dalam mengenai kesesuaian algoritma ML untuk prediksi kategori *BMI*.

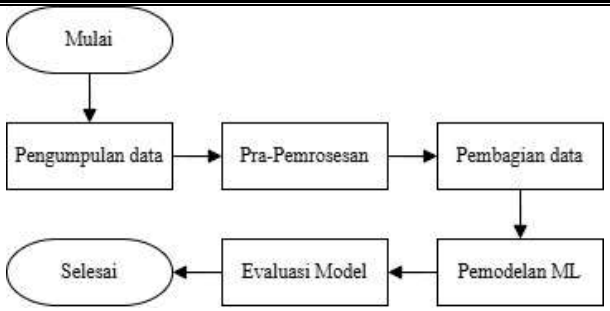
Penelitian ini bertujuan untuk menilai seberapa baik masing-masing kinerja algoritma, yaitu *Linear Regression*, *K-Nearest Neighbor*, dan *Gaussian Naive Bayes*, dalam memprediksi kategori *BMI* melalui penilaian secara menyeluruh yang mencakup akurasi klasifikasi serta kemampuan memperkirakan nilai *BMI* secara angka menggunakan *Mean Absolute Error (MAE)* dan *Root Mean Square Error (RMSE)*. Pendekatan ini memungkinkan analisis tidak hanya pada kemampuan klasifikasi kategori tetapi juga ketepatan perhitungan nilai *BMI* aktual. Dengan menilai semua ukuran ini secara menyeluruh, penelitian ini akan menentukan algoritma mana yang paling baik dan efisien untuk digunakan dalam sistem prediksi kategori *BMI* berdasarkan data yang dimasukkan pengguna.

2. METODOLOGI PENELITIAN

Studi ini bertujuan membandingkan kinerja tiga algoritma *Machine Learning* yang berbeda dalam memprediksi nilai *BMI* kontinu dan mengklasifikasikan kategori *BMI* berdasarkan data antropometri. Pendekatan penelitian termasuk dalam jenis analisis kuantitatif dengan rancangan komparatif eksperimental untuk menilai kemampuan model dalam mengelompokkan kategori *BMI* secara otomatis dan cepat seperti pada gambar 1.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **30/11/2025**; Accepted: **01/01/2026**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 1-6
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin



Gambar 1. Bagan Alur Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini bersumber dari repositori *Zenodo* dengan judul *BODY MASS INDEX CLASSIFICATION COMPARISON IN PREDICTING OBESITY AMONG ADULTS INDONESIAN USING K-MEANS CLUSTERING DATASET*. Dataset terdiri dari 111 baris data individu, dengan enam variabel yaitu Gender (jenis kelamin), Age (usia dalam tahun), Height (tinggi badan dalam meter), Weight (berat badan dalam kilogram), skor *BMI*, dan Label hasil perhitungan menggunakan rumus umum *BMI*.

$$\text{BMI} = \frac{\text{Weight(kg)}}{[\text{Height (m)}]^2} \quad (1)$$

Klasifikasi kategori *BMI* mengikuti standar *World Health Organization* (*WHO*) yang membagi *BMI* ke dalam empat kelompok seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Klasifikasi BMI Berdasarkan WHO

Rentang BMI (kg/m²)	Kategori
< 18.5	Underweight
18.5 – 24.9	Normal
25 – 29.9	Overweight
≥ 30	Obese

Untuk menggambarkan struktur data, Tabel 2 ditampilkan untuk menggambarkan sebagian isi dataset, yaitu beberapa baris awal dan akhir, dengan tanda "…" di tengah sebagai penanda baris yang tidak ditampilkan.

Tabel 2. Isi Dataset

No	Gender	Umur	Tinggi	Berat	BMI	Label
1	Perempuan	19	1,58	51	20,4294	Normal
2	Laki - laki	19	1,76	96,5	31,1532	Obese
3	Perempuan	19	1,58	47	18,8271	Normal
4	Laki - laki	19	1,76	49	15,8187	Normal
…	…	…	…	…	…	…
111	Perempuan	53	1,59	59	23,3377	Normal

2.2 Pra-pemrosesan Data

Tahap pra-pemrosesan data dilakukan melalui beberapa tahapan kunci. Pertama, variabel kategorikal Gender diubah menjadi bentuk numerik biner (Male = 0, Female = 1) sesuai dengan pendekatan yang telah digunakan dalam penelitian sejenis untuk data kesehatan [14]. Selanjutnya, seluruh fitur numerik distandardisasi menggunakan metode *z-score* untuk menyamakan skala data, mengikuti rekomendasi dalam studi pra-pemrosesan data kesehatan [15]. Proses standardisasi ini dilakukan secara manual dengan menghitung nilai rata-rata dan simpangan baku dari data latih. Pemeriksaan kualitas data mengonfirmasi tidak adanya data hilang maupun data ganda dalam dataset, sehingga memastikan kelayakan data untuk proses pemodelan selanjutnya.

2.3 Pembagian Data dan Evaluasi

Setelah melalui tahap pembersihan data, evaluasi kinerja model dilakukan menggunakan metode Stratified K-Fold Cross Validation [16]. Pendekatan ini dipilih untuk mengatasi keterbatasan jumlah sampel dengan memanfaatkan seluruh data secara optimal dalam proses pelatihan dan pengujian. Dataset dibagi secara acak menjadi lima bagian (fold) yang memiliki komposisi kategori *BMI* yang seimbang, di mana pada setiap iterasi empat bagian (80% data) digunakan untuk melatih model dan satu bagian sisanya (20% data) digunakan untuk menguji model [17]. Proses ini diulang secara sistematis

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **30/11/2025**; Accepted: **01/01/2026**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 1-6
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

hingga semua bagian telah bergiliran menjadi data uji. Hasil evaluasi dari setiap iterasi kemudian dirata-ratakan untuk memberikan estimasi kinerja model yang lebih andal dan stabil dibandingkan metode pembagian data tunggal [18]. Dengan demikian, model dapat diuji secara komprehensif pada berbagai subset data yang berbeda, sekaligus memastikan bahwa setiap kategori *BMI* terwakili secara proporsional dalam data latih dan uji.

2.4 Pemodelan ML

2.4.1 Linear Regression

Algoritma *Linear Regression* digunakan untuk memodelkan hubungan linear antara variabel input dengan variabel target kontinu. Model ini bekerja dengan menyesuaikan parameter θ yang meminimalkan error antara nilai prediksi dan nilai aktual. Parameter tersebut dihitung menggunakan persamaan normal.

$$\theta = \left(X^T X + \lambda I\right)^{-1} X^T y \quad (2)$$

di mana λ merupakan parameter regularisasi untuk menjaga stabilitas numerik. Pada penelitian ini, nilai λ digunakan dalam skala kecil dan tidak dilakukan pengujian atau penyesuaian nilai λ secara khusus, karena model Linear Regression digunakan sebagai model dasar tanpa optimasi parameter. Model linear sederhana terbukti cukup efektif untuk masalah prediksi nilai kesehatan [19]. Dalam implementasinya, langkah pertama adalah menambahkan komponen bias agar model dapat menyesuaikan garis prediksi dengan lebih fleksibel. Selanjutnya, parameter dihitung menggunakan pendekatan persamaan normal yang berfungsi untuk meminimalkan selisih antara nilai prediksi dan nilai sebenarnya. Dengan cara ini, model menghasilkan garis terbaik yang mewakili pola data.

2.4.2 Gaussian Naive Bayes

Algoritma *Gaussian Naive Bayes* merupakan metode klasifikasi probabilistik yang berdasarkan pada teorema Bayes dengan asumsi independensi antar fitur. Model ini menghitung probabilitas posterior suatu kelas diberikan fitur input menggunakan rumus umum *Gaussian Naive Bayes*.

$$P(c|x) \propto P(c) \times \prod P\left(x_j|c\right) \quad (3)$$

Di mana P(c) menunjukkan peluang awal suatu kelas dan P(xj|c) menunjukkan seberapa besar kemungkinan suatu fitur muncul pada kelas tersebut. Keunggulan algoritma ini terletak pada kesederhanaan komputasinya dan efektivitasnya ketika diterapkan untuk dataset dengan ukuran terbatas atau kecil hingga sedang [20]. Pada implementasi, program terlebih dahulu menghitung probabilitas awal tiap kelas, kemudian mencari rata-rata dan variasi dari fitur pada masing-masing kelas. Informasi ini digunakan untuk menilai seberapa besar kemungkinan data baru termasuk ke dalam kelas tertentu. Setelah peluang tiap kelas dihitung, hasilnya digabungkan untuk memperkirakan nilai *BMI* berdasarkan rata-rata *BMI* dari kelas yang relevan.

2.4.3 K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* bekerja dengan mencari sejumlah data latih terdekat dengan data coba berdasarkan kesamaan fitur. Tingkat kemiripan dihitung menggunakan jarak Euclidean.

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n}\left(x_j - x_{ij}\right)^2} \quad (4)$$

Setelah jarak antara data uji dan seluruh data latih dihitung, model akan menyusunnya dari yang paling dekat hingga yang paling jauh, lalu memilih sejumlah tetangga terdekat sesuai nilai K. Prediksi ditentukan dari rata-rata nilai tetangga tersebut. Penelitian oleh A'yuniyah dan Reza (2022) [21] menunjukkan bahwa penggunaan K=3 cukup efektif untuk data berukuran kecil hingga sedang, membuatnya relevan dengan dataset yang digunakan pada penelitian ini. Oleh karena itu, nilai K pada penelitian ini ditetapkan sebesar 3 dengan mempertimbangkan jumlah data yang digunakan serta mengacu pada hasil penelitian sebelumnya, dan tidak dilakukan pengujian atau penyesuaian nilai K secara khusus. Karakteristik tadi, dapat menangkap pola lokal dengan baik tanpa menimbulkan masalah kelebihan penyesuaian. Dalam program, langkah pertama adalah menghitung jarak antara data uji dengan seluruh data latih menggunakan rumus Euclidean. Jarak tersebut kemudian diurutkan dari yang paling dekat hingga paling jauh. Setelah itu, dipilih sejumlah tetangga terdekat sesuai nilai K, yaitu tiga tetangga dalam penelitian ini. Nilai prediksi ditentukan dengan mengambil rata-rata dari tetangga tersebut.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi dengan Stratified Cross Validation

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **30/11/2025**; Accepted: **01/01/2026**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 No. 1 Edisi Januari-Februari 2026, Page 1-6
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Pengujian dilakukan menggunakan *5-Fold Stratified Cross Validation* untuk menilai konsistensi kinerja algoritma. Tabel III menampilkan hasil MAE pada setiap fold.

Tabel 3. Hasil MAE per Fold dengan Stratified Cross Validation

Fold	Linear Regression	K-Nearest Neighbor	Gaussian Naive Bayes
1	0.457	1.573	2.855
2	0.281	2.302	3.208
3	0.383	1.166	2.574
4	0.433	1.655	2.254
5	0.256	1.222	2.256

Dari tabel terlihat bahwa *Linear Regression* konsisten dengan nilai MAE rendah di semua *fold* (0.25–0.45). *K-Nearest Neighbor* lebih fluktuatif, kadang mendekati *Linear Regression* (Fold 3: MAE 1.16), tetapi kadang jauh lebih tinggi (Fold 2: MAE 2.30). *Gaussian Naive Bayes* selalu memiliki MAE lebih besar (2.25–3.20), menandakan prediksi angka *BMI* kurang presisi.

3.2 Hasil Rata-Rata Kinerja Algoritma

Tabel 4. Hasil Rata-Rata Evaluasi Algoritma

Algoritma	MAE (kg/m²)	RMSE (kg/m²)	Akurasi Kategori (%)
Linear Regression	0.362	2.855	73.87
K-Nearest Neighbor	1.583	3.208	66.90
Gaussian Naive Bayes	2.629	2.574	75.74

Hasil ini menunjukkan adanya perbedaan fokus tiap algoritma:

a. *Linear Regression* paling tepat untuk menghitung angka *BMI* karena kesalahan rata-rata paling kecil.
b. *Gaussian Naive Bayes* unggul dalam klasifikasi kategori *BMI* dengan akurasi tertinggi.
c. *K-Nearest Neighbor* memiliki performa paling rendah pada dataset kecil, baik dalam prediksi angka maupun klasifikasi.

3.3 Validasi dengan Data Pengguna Baru

Untuk melihat penerapan nyata, sistem diuji pada data seorang laki-laki usia 21 tahun, tinggi 1,74 m, berat 54 kg. Nilai *BMI* sebenarnya adalah 17,84 (kategori Underweight). Hasil prediksi ditunjukkan pada Tabel 5.

Tabel 5. Hasil Prediksi BMI pada Data Pengguna Baru

Algoritma	Prediksi BMI	Error Absolut	Kategori WHO
Linear Regression	17.48	0.35	Underweight
K-Nearest Neighbor	17.60	0.23	Underweight
Gaussian Naive Bayes	22.22	4.38	Underweight

Hasil menunjukkan bahwa *K-Nearest Neighbor* paling mendekati nilai *BMI* aktual (error 0.23), dii-kuti oleh *Linear Regression* (error 0.35). *Gaussian Naive Bayes* meleset cukup jauh (error 4.38), tetapi tetap benar dalam menentukan kategori Underweight.

4. KESIMPULAN

Dari hasil penelitian ini dapat disimpulkan bahwa penerapan tiga algoritma Machine Learning, yaitu Linear Regression, K-Nearest Neighbor, dan Gaussian Naive Bayes, menunjukkan perbedaan karakteristik dalam memprediksi kategori Indeks Massa Tubuh (BMI) berdasarkan data dasar tubuh seperti tinggi, berat, usia, dan jenis kelamin. Gaussian Naive Bayes menunjukkan kinerja yang baik dalam klasifikasi kategori BMI, sedangkan Linear Regression lebih unggul dalam menghasilkan prediksi nilai BMI secara numerik. K-Nearest Neighbor menunjukkan performa yang lebih bervariasi pada dataset berukuran kecil, namun pada pengujian data pengguna baru mampu memberikan prediksi yang mendekati nilai aktual. Kelebihan penelitian ini terletak pada analisis menyeluruh terhadap klasifikasi dan prediksi numerik, serta penggunaan metode evaluasi yang konsisten. Namun, terdapat keterbatasan berupa ukuran dataset yang kecil dan belum adanya pengujian pada data yang lebih beragam. Ke depannya, pengembangan dapat dilakukan dengan memperluas jumlah dan jenis data, menambahkan fitur tambahan seperti aktivitas fisik atau pola makan, serta menguji

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **30/11/2025**; Accepted: **01/01/2026**; Published: **15/01/2026**

algoritma lain yang lebih kompleks untuk meningkatkan akurasi dan generalisasi model. Dengan pengembangan tersebut, sistem prediksi BMI berbasis Machine Learning berpotensi menjadi alat bantu yang efektif dalam mendukung layanan kesehatan primer dan platform digital berbasis data pengguna.

REFERENCES

[1] E. Asmin *et al.*, "PENYULUHAN PENYAKIT TIDAK MENULAR PADA MASYARAKAT," *Communnity Development Journal*, vol. 2, no. 3, pp. 940–944, 2021.

[2] D. Khanna, C. Peltzer, P. Kahar, and M. S. Parmar, "Body Mass Index (BMI): A Screening Tool Analysis," *Cureus*, Feb. 2022, doi: 10.7759/cureus.22119.

[3] Badan Kebijakan Pembangunan Kesehatan (BKPK), "Survei Kesehatan Indonesia (SKI) 2023 Dalam Angka," 2023.

[4] G. Al Raffi and N. Panji Purnomo, "PENGARUH UMUR, JENIS KELAMIN, MEROKOK, DAN NILAI BODY MASS INDEX (BMI) PADA RISIKO SESEORANG TERKENA STROKE MENGGUNAKAN REGRESI LOGISTIK," *Jurnal Inovasi Global*, vol. 3, no. 11, 2024, doi: 10.58344/jig.v2i11.

[5] Q. R. Cahyani *et al.*, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.

[6] J. Hosen, B. S. Engineering, and I. Ahmed, "Comparison of Regression, K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) Models for the Prediction of Weight, Gender and Body Mass Index Status," 2023.

[7] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.

[8] T. A. D. Kurniawan, A. Setiawan, and F. Tita, "Perbandingan Kinerja Metode Support Vector Regression dan Metode Regresi Linier Berganda dalam Memprediksi BMI pada Dataset ASTHMA," *Jurnal Sains dan Edukasi Sains*, vol. 8, no. 2, pp. 133–142, Aug. 2025, doi: 10.24246/juses.v8i2p133-142.

[9] I. Werdiningsih *et al.*, "Analisis Prediksi Stroke Menggunakan Pendekatan Decision Tree dengan Seleksi Fitur dan Neural Network," 2023.

[10] J. Lemantara and T. Lusiani, "ANALISIS PREDIKSI PENYAKIT DIABETES PADA WANITA MENGGUNAKAN METODE NAÏVE BAYES DAN K-NEAREST NEIGHBOR," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4911.

[11] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of risk factors associated with obesity and overweight—a machine learning overview," *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, doi: 10.3390/s20092734.

[12] B. A. Putra, N. Fadilah, H. Mukhtar, M. Fatchiyah Maharani, and A. Addarisalam, "Prediksi Risiko Depresi Pascapersalinan Menggunakan Algoritma K-Nearest Neighbor (KNN)," 2025.

[13] A. Fadli *et al.*, "PENGGUNAAN ALGORITMA NAIVE BAYES UNTUK MEMPREDIKSI KELULUSAN MAHASISWA," 2024. [Online]. Available: https://www.kaggle.com/datasets/hafizhathallah/kelul

[14] S. Abrori and Z. Fatah, "Implementasi Metode Decission Tree Dalam Mengklasifikasi Depresi Menggunakan Rapidminer," *Journal of Students' Research in Computer Science*, vol. 5, no. 2, pp. 123–132, Nov. 2024, doi: 10.31599/vgf7xb32.

[15] N. R. Febriyanti, K. Kusrini, and A. D. Hartanto, "Analisis Perbandingan Algoritma SVM, Random Forest dan Logistic Regression untuk Prediksi Stunting Balita," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 149–158, Apr. 2025, doi: 10.29408/edumatic.v9i1.29407.

[16] A. Oktaviana, D. P. Wijaya, A. Pramuntadi, and D. Heksaputra, "Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 812–818, May 2024, doi: 10.57152/malcom.v4i3.1268.

[17] Z. A. Sejuti and M. S. Islam, "A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation," *Sensors International*, vol. 4, Jan. 2023, doi: 10.1016/j.sintl.2023.100229.

[18] A. Desiani, D. A. Zayanti, I. Ramayanti, F. F. Ramadhan, and Giovillando, "PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN LOGISTIC REGRESSION DALAM KLASIFIKASI KANKER PAYUDARA," *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 4, no. 1, pp. 33–42, Jan. 2025, doi: 10.69916/jkbti.v4i1.191.

[19] K. Bartol, D. Bojanić, T. Petković, S. Peharec, and T. Pribanić, "Linear Regression vs. Deep Learning: A Simple Yet Effective Baseline for Human Body Measurement," *Sensors*, vol. 22, no. 5, Mar. 2022, doi: 10.3390/s22051885.

[20] A. M. Majid and I. Nawangsih, "Implementasi Machine Learning Menggunakan Adaboost dalam Prediksi Status Gizi Anak di Posyandu Tanjung XXIV," 2024.

[21] Q. A'yuniyah and M. Reza, "IJIRSE: Indonesian Journal of Informatic Research and Software Engineering Application Of The K-Nearest Neighbor Algorithm For Student Department Classification At 15 Pekanbaru State High School Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," 2022.