Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

Comparative Evaluation of Linear Regression and Ensemble Learning Models for Daily Calorie Prediction Using a Public Lifestyle Dataset with Structured Preprocessing and Recursive Feature Elimination

Yunandra Wahyu Utama*[1], Majid Rahardi[2]

[1,2]Informatics, Universitas AMIKOM Yogyakarta, Indonesia

Email: [1]yunandrawu03@gmail.com

Received: Jan 6, 2026; Revised: Mar 22, 2026; Accepted: Mar 10, 2026; Published: Jun 15, 2026

Abstract

Accurate daily calorie estimates are essential for personalized nutrition and prevention of diet-related conditions, yet lifestyle variability can reduce the effectiveness of one-size-fits-all recommendations. This study aims to develop an accurate lifestyle-based calorie estimation model by comparing an interpretable linear approach with ensemble machine learning methods. A publicly available lifestyle dataset from Kaggle was used, containing demographic variables, anthropometric measurements, food intake, dietary patterns, and physical activity attributes. A preprocessing pipeline was applied, including outlier handling using interquartile range capping, categorical encoding, normalization, and feature selection via Recursive Feature Elimination to identify the most relevant predictors. Four models (Linear Regression, Random Forest, XGBoost, and LightGBM) were trained and evaluated, followed by hyperparameter tuning of ensemble models using GridSearchCV. Performance was assessed using R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) and training time. Linear Regression achieved the best overall performance (R^2 = 0.9650, MAE = 80.95, RMSE = 101.71, training time = 8.95 seconds). Among ensembles, the tuned XGBoost performed best (R^2 = 0.9646, MAE = 81.34, RMSE = 102.35, training time = 10.55 seconds). Compared with tuned XGBoost, Linear Regression was superior with MAE by 0.39 and RMSE by 0.64, while R^2 increased by 0.0004 and required less computational time, indicating that added complexity did not yield meaningful gains on this structured dataset. These findings suggest that, for structured lifestyle data, interpretable linear models can match or outperform complex ensembles while remaining computationally efficient for real-time or edge-deployed health applications.

Keywords : *Calorie Estimation, Feature Selection, Lifestyle Data, Regression Analysis, Supervised Learning.*

1. INTRODUCTION

Kalori merupakan satuan energi yang diperoleh dari makanan, yang digunakan tubuh untuk menjalankan metabolisme, aktivitas fisik, serta mempertahankan proses fisiologis dasar [1]. Namun, ketidakseimbangan antara asupan dan penggunaan energi masih menjadi persoalan gizi yang menonjol. Berdasarkan Survei Kesehatan Indonesia (SKI) pada tahun 2023 yang dirilis oleh Kementerian Kesehatan RI, prevalensi obesitas dewasa meningkat dari 21,8% (Riskesdas 2018) menjadi 23,4% pada tahun 2023. Hal itu menunjukkan adanya tren peningkatan signifikan yang dipengaruhi oleh pola hidup dan perilaku makan masyarakat. Aktivitas fisik yang rendah, tingginya konsumsi energi, serta kebiasaan makan yang tidak teratur berkaitan dengan meningkatnya risiko overweight pada remaja [2]. Status gizi juga terbukti sangat dipengaruhi oleh asupan karbohidrat, protein, lemak, dan total energi yang dikonsumsi, sehingga ketidakseimbangan gizi makro menjadi faktor penentu penting dalam kondisi gizi remaja dan dewasa muda [3]. Di sisi yang lain, beberapa kelompok justru mengalami defisit energi meskipun angka gizi lebih tetap tinggi, menggambarkan ketidaksesuaian antara kebutuhan metabolik dan pola konsumsi harian [4]. Defisit energi dan makronutrien dalam skala besar bahkan ditemukan pada berbagai populasi remaja, menandakan rendahnya pemenuhan kebutuhan energi harian [5]. Selain

itu, pengendalian berat badan dapat dilakukan dengan membatasi total kalori harian sesuai target untuk mencegah peningkatan BMI dan menurunkan risiko obesitas [6].

Perkembangan modernisasi juga mendorong perubahan signifikan pada pola makan dan aktivitas masyarakat, seiring meningkatnya ketergantungan terhadap teknologi digital dalam kehidupan sehari-hari [7]. Perubahan ini juga menyebabkan banyak individu mengalami kesulitan dalam mengelola keseimbangan energi, termasuk dalam menghitung kebutuhan kalori harian yang secara manual sering tidak akurat karena memerlukan pengetahuan gizi, perhitungan detail, serta pemantauan konsumsi yang konsisten, sehingga rentan terjadi kesalahan dan sulit diterapkan secara berkelanjutan [8]. Penentuan kebutuhan kalori harian pun masih sering dilakukan secara manual (misalnya menggunakan rumus tertentu), sehingga diperlukan sistem yang membuat proses perhitungan lebih mudah dan efektif bagi pengguna [9]. Melihat tantangan tersebut, teknologi digital mulai dimanfaatkan sebagai alat bantu dalam pengolahan informasi kesehatan, di mana sistem komputasi mampu melakukan perhitungan otomatis dan menyediakan rekomendasi berbasis data secara lebih efisien [8]. Pendekatan berbasis aplikasi mobile juga dikembangkan untuk membantu pencatatan dan pengendalian asupan kalori harian, sekaligus dievaluasi dari sisi penerimaan pengguna [10]. Pemanfaatan artificial intelligence (AI) untuk mengukur asupan makanan dan nutrien juga terus berkembang karena berpotensi meningkatkan objektivitas, mengurangi recall bias, dan mendukung pemantauan real-time pada asesmen diet [11]. Sistem rekomendasi diet yang umum masih cenderung berbasis pedoman umum dan kurang adaptif terhadap perubahan gaya hidup individu, sehingga personalisasi prediksi kebutuhan energi menjadi kebutuhan penting [12]. Dalam konteks estimasi kebutuhan energi, indirect calorimetry dikenal sebagai gold standard untuk mengukur resting energy expenditure (REE), tetapi penerapannya sering terbatas karena kompleks dan mahal, sehingga pendekatan machine learning dinilai dapat meningkatkan akurasi estimasi [13]. Perkembangan teknologi lebih lanjut ditunjukkan oleh machine learning, yang memiliki kemampuan untuk mengenali pola kompleks dalam data kesehatan, mengekstraksi hubungan antara gaya hidup, nutrisi, dan status metabolik, serta meningkatkan akurasi prediksi dibandingkan metode konvensional [14]. Bahkan, penerapannya dalam pemodelan kesehatan telah terbukti efektif dalam menghasilkan prediksi yang lebih presisi, termasuk estimasi kebutuhan energi harian berdasarkan faktor gaya hidup multidimensi [15]. Di sisi lain, pengukuran konsumsi atau pengeluaran energi dapat dipengaruhi variasi individu serta perbedaan metode (misalnya ergocycle dan accelerometer), sehingga pemilihan pendekatan yang tepat menjadi aspek penting untuk memperoleh prediksi yang akurat [16].

Penelitian mengenai prediksi kalori dan pemodelan kesehatan berbasis machine learning telah berkembang pesat dalam beberapa tahun terakhir. Prediksi kalori berbasis data aktivitas fisik menggunakan XGBoost yang dioptimasi dengan Bayesian Optimization dan Nested Cross Validation menghasilkan MSE 4294.27, RMSE 65.53, dan R^2 0.9917, yang jauh melampaui model baseline seperti Random Forest maupun SVM [17]. Evaluasi efektivitas algoritma AdaBoost dan XGBoost dalam prediksi obesitas pada populasi dewasa menghasilkan akurasi, presisi, dan recall sebesar 92%, sedangkan AdaBoost hanya mencapai akurasi 40%, presisi 36%, dan recall 40% [18]. Pada klasifikasi risiko kesehatan makanan menggunakan indikator seperti glikemik indeks dan proporsi makronutrien, Multiple Linear Regression (MLR) memiliki performa paling unggul dalam estimasi energi, dengan skor $R^2 = 0.99$, MAE 7.7 kcal dan RMSE 18 kcal pada data training maupun testing, menandakan tingkat akurasi yang sangat tinggi dalam memprediksi kalori makanan dibandingkan model Random Forest Regression dan Decision Tree [19]. Pada penelitian mengenai prediksi kepadatan energi menggunakan Multiple Linear Regression menyimpulkan bahwa model yang sederhana sudah sangat memadai dan akurat dengan nilai R^2 mencapai 0.869 [20]. Sistem prediksi kalori menggunakan deep learning (CNN-LSTM dan Transformer) menunjukkan performa prediksi yang sangat tinggi dengan akurasi mencapai 93.8% pada model CNN-LSTM dan 95.1% pada model Transformer-based Deep Learning, khususnya pada data multimodal gabungan sensor wearable dan input makanan pengguna [21]. Pada pemodelan

kalori terbakar melalui regresi non-linear, XGBoost Regressor menghasilkan MAE sebesar 2.71, jauh lebih rendah dibandingkan algoritma baseline lainnya [22]. Hybrid model analisis untuk prediksi kalori dengan membandingkan performa dua model ensemble, yaitu Random Forest dan XGBoost. XGBoost memiliki performa prediksi yang lebih baik, dengan nilai R^2 dan error (MAE dan RMSE) lebih rendah dibandingkan Random Forest [23]. Estimasi kalori juga dapat dilakukan melalui pengolahan citra makanan pada perangkat mobile menggunakan kombinasi CNN–YOLO untuk deteksi real-time [24]. Sistem pengenalan makanan berbasis deep learning juga dapat menunjukkan bahwa identifikasi jenis makanan dapat diintegrasikan dengan estimasi kalori untuk membantu pemantauan asupan dan pencegahan obesitas secara lebih praktis [25]. Namun, pendekatan berbasis citra 2D sering menghadapi keterbatasan dalam memperkirakan porsi karena tidak menangkap informasi volume, sehingga metode rekonstruksi 3D mulai banyak diteliti untuk meningkatkan presisi estimasi kalori [26]. Inovasi terbaru juga mengarah pada integrasi machine learning dengan Large Language Models (LLM) untuk menghasilkan rekomendasi diet awal yang lebih personal berbasis hasil prediksi [27]. Kompleksitas data gaya hidup dan aktivitas juga menuntut rancangan tahap preprocessing serta optimasi model agar prediksi kalori lebih stabil dan konsisten pada berbagai kondisi pengguna [28].

Meskipun sejumlah penelitian telah memanfaatkan machine learning dalam prediksi kalori dan pemodelan kesehatan, masih ada kesenjangan penelitian yang layak dieksplorasi dalam penelitian ini. Beberapa studi telah menerapkan optimasi menggunakan Bayesian Optimization dan Nested Cross Validation, namun perbandingan antara model linear dan beberapa model ensemble dengan prosedur optimasi yang setara masih belum banyak dilaporkan. Kondisi ini menunjukkan perlunya rancangan eksperimen yang terstruktur untuk menilai model prediksi kalori harian berbasis variabel gaya hidup secara lebih objektif. Namun, penelitian yang membandingkan secara komprehensif model baseline linear dan beberapa model ensemble pada prediksi kebutuhan kalori harian berbasis gaya hidup dengan tahapan praproses yang sama serta strategi optimasi hyperparameter yang setara masih terbatas, sehingga sulit memastikan apakah peningkatan kinerja benar-benar berasal dari algoritma atau dari perbedaan langkah pengolahan data dan metode tuning. Secara umum, literatur terkini menunjukkan bahwa penelitian prediksi kalori banyak menitikberatkan peningkatan akurasi melalui model ensemble dan deep learning, baik pada data aktivitas fisik maupun pendekatan berbasis citra, dengan capaian kinerja yang tinggi pada berbagai metrik evaluasi [17]–[19], [21]–[26]. Namun, pendekatan berbasis citra dan deep learning umumnya menuntut komputasi lebih tinggi serta sensitif terhadap kualitas input dan estimasi porsi, sehingga tidak selalu praktis untuk implementasi yang ringan pada perangkat pengguna [24]–[26]. Sementara itu, studi yang mengeksplor model regresi yang lebih ringan berbasis atribut gaya hidup non-invasif untuk memprediksi kebutuhan kalori harian masih relatif terbatas dan belum banyak dibandingkan secara komprehensif dalam rancangan evaluasi yang konsisten. Hal ini sejalan dengan studi perbandingan pada domain kesehatan yang menunjukkan bahwa variasi rancangan studi, pemilihan prediktor, metode validasi, serta strategi tuning dapat menyebabkan perbedaan performa antar model, sehingga evaluasi yang terstandar diperlukan agar perbandingan benar-benar adil [29], [30]. Berdasarkan gap tersebut, novelty penelitian ini terletak pada penyusunan rancangan evaluasi yang terstandar dan adil untuk membandingkan model linear dan beberapa model ensemble dalam prediksi kebutuhan kalori harian berbasis gaya hidup, dengan tahapan praproses data serta strategi optimasi yang dibuat setara sehingga perbedaan kinerja yang diperoleh benar-benar mencerminkan kontribusi algoritma.

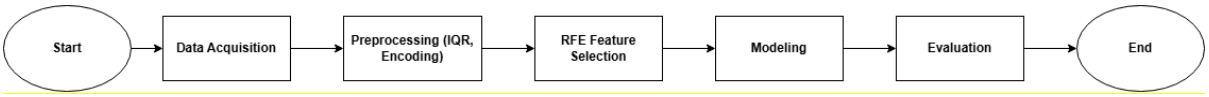
Berdasarkan celah tersebut, penelitian ini bertujuan mengembangkan model prediksi kalori harian berbasis gaya hidup menggunakan dataset publik yang berasal dari Kaggle dengan menerapkan tahapan praproses data preprocessing yang terstruktur, meliputi penanganan outlier menggunakan IQR capping, encoding variabel kategorikal, normalisasi, serta seleksi fitur menggunakan Recursive Feature Elimination (RFE). Penelitian ini mengevaluasi Linear Regression sebagai baseline, serta model

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

ensemble (Random Forest, XGBoost, dan LightGBM) dioptimasi menggunakan GridSearchCV. Kinerja setiap model dievaluasi menggunakan metrik R^2, Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE), sehingga menghasilkan rekomendasi algoritma yang paling optimal untuk prediksi kebutuhan kalori harian berbasis gaya hidup.

2. METHOD

Metodologi yang digunakan dalam penelitian ini disusun secara sistematis untuk memastikan proses analisis data dan pembangunan model berjalan secara terstruktur. Secara garis besar, tahapan penelitian dimulai dari identifikasi dataset publik, prapemrosesan, seleksi fitur, hingga tahap evaluasi, yang secara visual direpresentasikan dalam Gambar 1 berikut.



Gambar 1. Alur Penelitian

Gambar 1 menunjukkan alur penelitian yang dimulai dengan tahap Data Acquisition untuk mengambil dataset sekunder yang bersumber dari platform Kaggle. Tahap selanjutnya adalah preprocessing yang mencakup penanganan outlier menggunakan metode IQR serta melakukan encoding pada variable kategorikal agar data siap diproses oleh algoritma. Setelah data bersih, dilakukan RFE Feature Selection untuk memilih fitur-fitur yang relevan guna meningkatkan efisiensi model. Tahap modeling melibatkan penggunaan empat algoritma, yaitu Linear Regression, Random Forest, XGBoost, dan LightGBM. Seluruh model kemudian masuk ke tahap evaluasi menggunakan metrik R^2, MAE, dan RMSE untuk menentukan performa model yang telah dibangun.

2.1. Data Acquisition

Penelitian ini menggunakan dataset publik "Life Style Data" yang tersedia pada platform Kaggle. Dataset yang digunakan berisi data gaya hidup yang mencakup variabel demografi, pengukuran antropometrik, serta informasi yang mempresentasikan pola konsumsi/asumsi makanan, pola makan, dan aktivitas fisik. Variabel target dalam penelitian ini adalah Calories, sehingga dataset dapat digunakan sebagai dasar pemodelan untuk memprediksi kebutuhan kalori harian berdasarkan karakteristik gaya hidup. Proses akuisisi data dilakukan dengan mengunduh dataset dari Kaggle dan menyimpan file Final_data.csv pada Google Drive agar dapat diakses melalui Google Colab. Google Drive terlebih dahulu di-mount menggunakan fungsi drive.mount('/content/drive'), kemudian direktori kerja diarahkan ke lokasi penyimpanan dataset, yaitu /content/drive/MyDrive/Jurnal/Dataset. Setelah itu, dataset dibaca menggunakan pustaka pandas melalui perintah pd.read_csv('Final_data.csv') dan disimpan ke dalam variabel df_raw. Untuk memastikan data berhasil dimuat dengan benar serta memverifikasi struktur awal dataset, dilakukan pengecekan dengan menampilkan beberapa baris pertama dan terakhir menggunakan df_raw.head(5) dan df_raw.tail(5) sebelum proses analisis eksploratif dan preprocessing data dilakukan pada tahap berikutnya.

2.2. Preprocessing Data

Preprocessing data dilakukan untuk menyiapkan dan menyesuaikan data dengan membersihkannya agar terlihat lebih rapi, terstruktur, serta lebih efisien dan efektif untuk dianalisis lebih lanjut [31]. Tahap pertama dalam preprocessing ini adalah memisahkan fitur menjadi numerik dan kategorikal dengan mengidentifikasi kolom bertipe numerik serta menentukan kolom kategorikal (Gender, meal_type, diet_type, dan cooking_method). Variabel kategorikal kemudian dikonversi menjadi bentuk numerik menggunakan LabelEncoder, sehingga seluruh fitur dapat diproses oleh

algoritma pembelajaran mesin. Setelah variabel kategorikal dikonversi ke bentuk numerik, tahap berikutnya adalah memastikan kualitas data numerik dengan mengidentifikasi outlier. Outlier adalah nilai yang menyimpang jauh dari sebagian besar data, yang dapat muncul akibat kesalahan pencatatan/pengukuran atau merepresentasikan kondisi khusus [32]. Keberadaan outlier pada fitur numerik diperiksa melalui visualisasi boxplot, lalu dilakukan penanganan outlier menggunakan metode IQR capping, yaitu membatasi nilai yang berada di bawah batas bawah atau di atas batas atas agar tidak mendistorsi proses pelatihan model. Setelah proses capping, boxplot kembali ditampilkan untuk memastikan sebaran data menjadi lebih stabil. Tahap berikutnya adalah normalisasi/standarisasi fitur menggunakan StandardScaler untuk menyamakan skala antar variabel, sehingga fitur dengan rentang besar tidak mendominasi pembelajaran model. Setelah seluruh proses preprocessing selesai, dataset akhir disiapkan dengan memisahkan fitur prediktor (X) dan variabel target (y), di mana target yang digunakan adalah Calories, sedangkan X merupakan seluruh fitur selain target.

2.3. Feature Selection (RFE)

Pada tahap seleksi fitur lanjutan, penelitian ini menggunakan Recursive Feature Elimination (RFE) untuk memilih fitur relevan sebelum pelatihan model. Recursive Feature Elimination (RFE) adalah teknik pemilihan fitur yang menjalankan pelatihan model lalu mengurutkan dan menghapus fitur dengan kontribusi terendah secara bertahap [33]. Metode ini banyak digunakan karena mampu mengelola data berukuran besar dengan lebih efisien serta membantu meningkatkan performa prediksi dan kejelasan interpretasi model [34]. Empat algoritma yang digunakan, yaitu Linear Regression, Random Forest, XGBoost, dan LightGBM, masing-masing dipasangkan dengan RFE agar proses seleksi fitur mempertimbangkan karakteristik model yang berbeda. Dalam implementasinya, jumlah fitur yang dipilih ditetapkan sebanyak 15 fitur (n_features_to_select = 15). Setelah RFE menghasilkan himpunan fitur terpilih untuk setiap model, data latih dan data uji kemudian diproyeksikan hanya pada fitur tersebut, sehingga pelatihan dan pengujian model dilakukan pada fitur yang lebih ringkas dan informatif. Daftar fitur terpilih dari masing-masing model selanjutnya dicatat sebagai keluaran seleksi fitur untuk digunakan pada tahap evaluasi dan optimasi model berikutnya.

2.4. Modeling

Dalam tahap ini, dilakukan pelatihan empat algoritma, yaitu Linear Regression, Random Forest, XGBoost, dan LightGBM. Dasar pengembangan model ini berfokus pada minimalisasi fungsi objektif masing-masing algoritma. Pada Linear Regression, fungsi objektif bertujuan meminimalkan Residual Sum of Squares (RSS) sebagaimana dinyatakan dalam persamaan 1.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

Sedangkan pada XGBoost, fungsi objektif (Obj) menggabungkan Loss Function (L) dan Regularization (**Ω**) untuk mengontrol kompleksitas model guna mencegah overfitting, sesuai dengan persamaan (2).

$$Obj = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (2)$$

Setelah diperoleh fitur terpilih sebanyak 15 fitur melalui RFE masing-masing model, data latih (X_train) dan data uji (X_test) dibatasi hanya pada fitur-fitur tersebut. Selanjutnya, setiap model dilatih menggunakan data latih terpilih (model.fit(X_train_selected, y_train)), lalu digunakan untuk menghasilkan prediksi pada data uji (model.predict(X_test_selected)). Dengan demikian, proses pelatihan dilakukan pada subset fitur yang paling relevan untuk masing-masing algoritma agar pembelajaran model lebih efektif dan hasil evaluasi antar model dapat dibandingkan dalam kerangka yang konsisten.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

Pada tahap peningkatan performa model menerapkan hyperparameter tuning menggunakan teknik GridSearchCV. GridSearchCV merupakan teknik optimasi hyperparameter yang mengevaluasi berbagai kombinasi parameter melalui proses cross-validation otomatis guna memperoleh model yang paling optimal tanpa harus memeriksanya secara manual [35]. Ruang pencarian hiperparameter ditetapkan berbeda untuk tiap model, Linear Regression diuji pada variasi fit_intercept dan positive, Random Forest pada n_estimators dan max_depth, sedangkan XGBoost dan LightGBM pada n_estimators, max_depth, dan learning_rate. Rincian retang parameter yang diuji serta parameter terbaik yang dihasilkan oleh GridSearchCV disajikan pada Tabel 1.

Tabel 1 Parameter Grid dan Hasil Best Parameters GridSearchCV

Model	Hyperparameter	Parameter Grid	Best Parameter
Linear Regression	fit_intercept	[True, False]	True
Random Forest	n_estimators	[50, 100]	100
	max_depth	[10, 20]	20
	min_samples_leaf	[1, 2]	2
	max_features	[‘sqrt’]	‘sqrt’
XGBoost	n_estimators	[100, 200]	100
	max_depth	[3, 5]	3
	learning_rate	[0.01, 0.1]	0.1
LightGBM	n_estimators	[100, 200]	100
	num_leaves	[31, 63]	31
	learning_rate	[0.01, 0.1]	0.1

Untuk efisiensi komputasi, proses tuning dijalankan secara paralel menggunakan n_jobs=-1. Setelah proses pencarian selesai, model terbaik dipilih berdasarkan nilai R^2 tertinggi pada validasi silang (best_estimator_ dan best_params_). Model terbaik tersebut kemudian digunakan untuk memprediksi data uji, dan kinerjanya dihitung kembali menggunakan metrik R^2, MAE, dan RMSE.

2.5. Evaluation

Setelah model dilatih, performa setiap model dievaluasi menggunakan tiga metrik regresi, yaitu R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). Setelah tahap modeling dilakukan evaluasi dengan menghitung R^2 Score dengan rumus matematis yang dinyatakan dalam persamaan (3).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

R^2 (coefficient of determination) mengukur seberapa besar variasi pada variabel target yang dapat dijelaskan oleh model, dengan nilai terbaik 1, berada sekitar 0 bila setara dengan prediksi rata-rata, dan bisa menjadi negatif apabila kinerja model lebih buruk dibandingkan baseline tersebut [36]. Setelah R^2, model dievaluasi dengan menghitung RMSE berdasarkan rumus matematis pada persamaan (4).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

RMSE adalah akar dari rata-rata kuadrat selisih antara nilai aktual dan prediksi, sehingga kesalahan yang besar akan dihukum lebih berat, oleh karena itu semakin kecil nilai RMSE, semakin

baik kinerja model [37]. Dan yang terakhir, model dievaluasi dengan menghitung MAE berdasarkan rumus matematis pada persamaan (5).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (5)$$

MAE merupakan rata-rata selisih absolut antara nilai prediksi dan nilai aktual, sehingga dapat dipandang sebagai "rata-rata kesalahan" dalam satuan asli data dan lebih mudah diinterpretasikan secara langsung [38].

Nilai ketiga metrik dihitung dari hasil prediksi pada data uji, ditampilkan untuk setiap model, kemudian dirangkum dalam bentuk table menggunakan DataFrame agar perbandingan kinerja antar model dapat dilakukan secara terstruktur dan mudah dianalisis.

3. RESULT

3.1. Data Acquisition

Dataset memiliki 20.000 data dan 54 variabel, terdiri dari 39 fitur numerik (float64) dan 15 fitur kategorikal (object). Pemeriksaan kelengkapan data melalui df_raw.isna().sum() dan df_raw.isnull().sum() menunjukkan seluruh variabel memiliki nilai 0 untuk missing value, sehingga setiap kolom berisi 20.000 data non-null. Selain itu, df_raw.nunique() digunakan untuk melihat jumlah nilai unik pada tiap variabel untuk membantu mengidentifikasi keragaman nilai pada fitur numerik maupun jumlah kategori pada fitur kategorikal.

3.2. Preprocessing Data

3.2.1. Encoding Kategorikal

Tahap awal preprocessing data dilakukan dengan memisahkan fitur menjadi numerik dan kategorikal. Fitur numerik diidentifikasi menggunakan df.select_dtypes(include=[np.number]), sedangkan fitur kategorikal ditetapkan secara eksplisit yaitu Gender, meal_type, diet_type, dan cooking_method. Karena algoritma pembelajaran mesin pada penelitian ini bekerja pada representasi numerik, maka fitur kategorikal perlu diubah menjadi bilangan. Konversi dilakukan menggunakan LabelEncoder, yaitu mengubah setiap kategori menjadi kode bilangan bulat berdasarkan urutan kelas yang terdeteksi pada data. Hasil encoding (berdasarkan output print(encoder.classes_)) menunjukkan kategori pada masing-masing variabel , yaitu Gender (Female, Male), meal_type (Breakfast, Dinner, Lunch, Snack), diet_type (Balanced, Keto, Low-Carb, Paleo, Vegan, Vegetarian), dan cooking_method (Baked, Boiled, Fried, Grilled, Raw, Roasted, Steamed) bisa dilihat pada Tabel 2.

Tabel 2. Hasil Encoding Fitur Kategorikal

	Gender	meal_type	diet_type	cooking_method
0	1	2	4	3
1	0	2	5	2
2	0	0	3	1
3	0	2	3	2
4	1	0	4	0

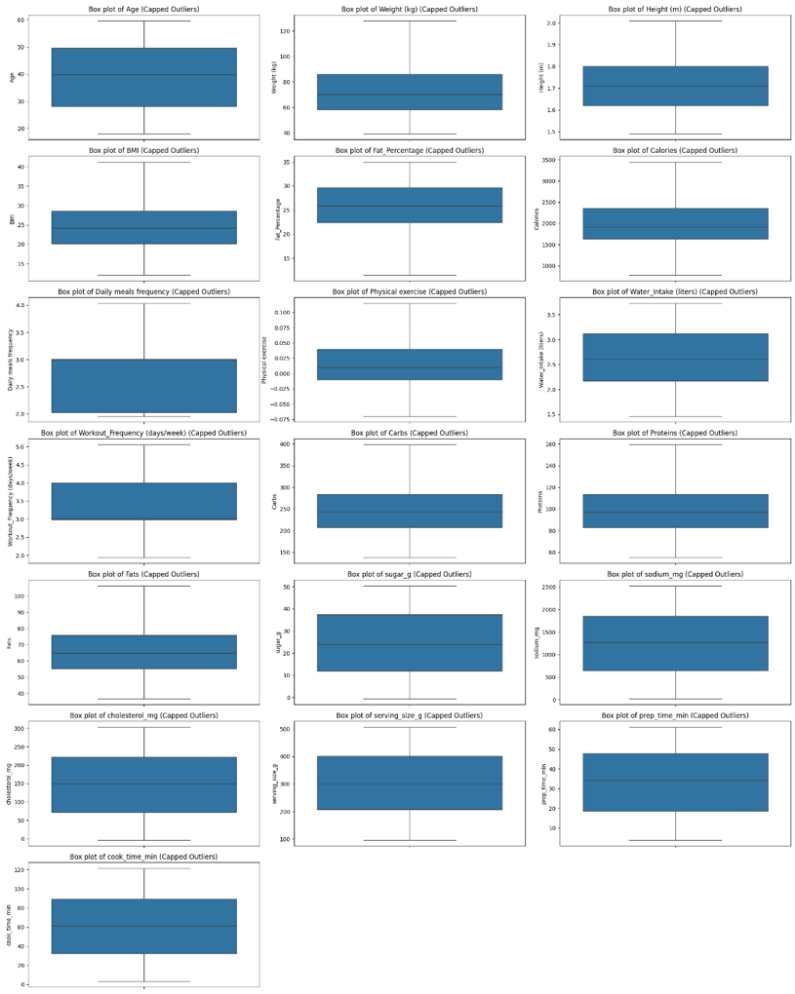
Pada Tabel 2. Hasil encoding menunjukkan daftar kategori yang terdeteksi pada masing-masing variabel yang terlihat pada Tabel 1, nilai kategorikal berubah menjadi kode bilangan bulat, dengan Gender (Female : 1, Male : 0), meal_type (Breakfast : 0, Dinner : 1, Lunch : 2, Snack : 3), diet_type

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

(Balanced : 0, Keto : 1, Low-Carb : 2, Paleo : 3, Vegan : 4, Vegetarian : 5), cooking_method (Baked : 0, Boiled : 1, Fried : 2, Grilled : 3, Raw : 4, Roasted : 5, Steamed : 6).

3.2.2. Deteksi dan Penanganan Outlier

Setelah variabel kategorikal di-encoding, langkah berikutnya adalah mendeteksi outlier pada fitur numerik. Outlier diperiksa menggunakan boxplot untuk seluruh fitur numerik (sns.boxplot pada setiap kolom). Pada boxplot awal terlihat adanya titik-titik data yang berada di luar whisker pada beberapa fitur, misalnya Weight (kg), BMI, Calories, serta beberapa fitur makronutrien seperti Carbs, Proteins, dan Fats. Kondisi ini menunjukkan adanya nilai ekstrem yang berpotensi memengaruhi proses pembelajaran model, khususnya model yang sensitif terhadap nilai ekstrem seperti regresi linear. Penanganan outlier dilakukan menggunakan metode IQR capping melalui fungsi cap_outliers_iqr(). Pada metode ini dihitung kuartil pertama (Q1), kuartil ketiga (Q3), dan Interquartile Range (IQR = Q3 − Q1), kemudian ditentukan batas bawah dan batas atas, dengan lower bound = Q1 − 1,5 × IQR dan upper bound = Q3 + 1,5 × IQR. Nilai yang berada di bawah batas bawah akan diganti menjadi lower bound, dan nilai yang berada di atas batas atas akan diganti menjadi upper bound. Proses ini diterapkan pada seluruh fitur numerik (for col in numerical_features). Setelah dilakukan deteksi outlier, dilakukan penanganan outlier yang bisa dilihat pada Gambar 2.



Gambar 2. Penanganan Outlier

Setelah dilakukan penanganan outlier menggunakan metode IQR capping pada Gambar 2, boxplot pada seluruh fitur numerik menunjukkan bahwa nilai-nilai yang sebelumnya berada jauh di luar

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

whisker (titik outlier) sudah tidak tampak menonjol lagi karena dibatasi pada batas bawah dan batas atas. Dampaknya, sebaran data menjadi lebih stabil (rentang whisker terlihat lebih representatif terhadap mayoritas data), sementara median dan rentang antar kuartil (IQR) pada sebagian besar variabel tetap relatif konsisten, artinya capping tidak mengubah pusat distribusi secara drastis tetapi hanya mengurangi pengaruh nilai ekstrem. Perubahan ini terutama bermanfaat pada fitur yang sebelumnya memiliki outlier jelas, seperti Weight (kg), BMI, Calories, serta variabel makronutrien (Carbs, Proteins, Fats) dan beberapa variabel gizi lain (misalnya sodium_mg). Dengan outlier yang sudah terkendali, risiko distorsi pada proses pelatihan model berkurang, terutama untuk model yang sensitif terhadap nilai ekstrem seperti Linear Regression, sehingga pembelajaran model menjadi lebih robust dan evaluasi kinerja menjadi lebih adil. Selain itu, hasil capping ini juga membuat tahap normalisasi/standarisasi berikutnya lebih efektif karena skala fitur tidak lagi dipengaruhi oleh nilai yang terlalu jauh dari pola umum data.

Setelah penanganan outlier dengan IQR capping untuk menstabilkan distribusi, dilakukan standarisasi menggunakan StandardScaler agar fitur berada pada skala sebanding, yang kemudian dilanjutkan dengan pembagian data menjadi data latih sebanyak 16.000 data (80%) dan data uji sebanyak 4.000 data (20%) untuk memastikan objektivitas evaluasi model.

3.3. RFE Feature Selection

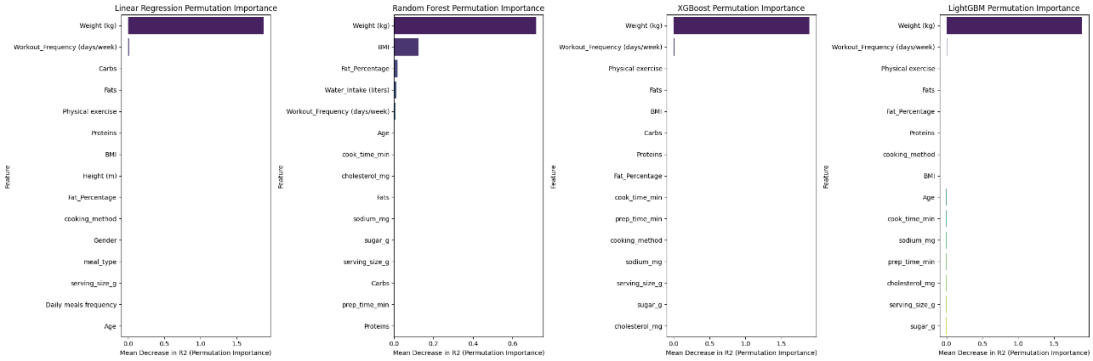
Proses seleksi fitur dilakukan menggunakan metode Recursive Feature Elimination (RFE) untuk mengidentifikasi variabel yang paling berpengaruh terhadap model. Hasil dari seleksi tersebut, yang merangkum 15 fitur terbaik untuk masing-masing algoritma, disajikan pada Tabel 3 di bawah ini.

Tabel 3 Hasil Seleksi Fitur RFE

Model	15 Fitur Terpilih
Linear Regression	Age, Gender, Weight (kg), Height (m), BMI, Fat_Percentage, Daily meals frequency, Physical exercise, Workout_Frequency (days/week), Carbs, Proteins, Fats, meal_type, serving_size_g, cooking_method
Random Forest	Age, Weight (kg), BMI, Fat_Percentage, Water_Intake (liters), Workout_Frequency (days/week), Carbs, Proteins, Fats, sugar_g, sodium_mg, cholesterol_mg, serving_size_g, prep_time_min, cook_time_min
XGBoost	Weight (kg), BMI, Fat_Percentage, Physical exercise, Workout_Frequency (days/week), Carbs, Proteins, Fats, sugar_g, sodium_mg, cholesterol_mg, serving_size_g, cooking_method, prep_time_min, cook_time_min
LightGBM	Age, Weight (kg), BMI, Fat_Percentage, Physical exercise, Workout_Frequency (days/week), Proteins, Fats, sugar_g, sodium_mg, cholesterol_mg, serving_size_g, cooking_method, prep_time_min, cook_time_min

Pada Tabel 3, hasil tahap seleksi fitur menggunakan Recursive Feature Elimination (RFE) menghasilkan 15 fitur terbaik yang spesifik untuk setiap algoritma. Secara umum, model Linear Regression memilih kombinasi atribut demografi, antropometrik, aktivitas, dan nutrisi. Sementara itu, model ensemble (Random Forest, XGBoost, LightGBM) lebih banyak menekankan pada variabel tubuh, aktivitas fisik, makronutrien, dan komponen gizi spesifik lainnya. Penggunaan subset fitur ini bertujuan untuk menekan risiko overfitting serta meningkatkan efisiensi komputasi tanpa mengorbankan akurasi prediksi. Untuk memberikan gambaran yang lebih mendalam mengenai kontribusi fitur terhadap prediksi, hasil permutation importance disajikan pada Gambar 3.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621



Gambar 3. Permutation Importance

Berdasarkan analisis pada Gambar 3, menunjukkan karakteristik unik dari setiap model dalam menangkap pengaruh fitur terhadap prediksi kalori. Pada Linear Regression, fitur Weight (kg) mendominasi secara mutlak sebagai prediktor utama dengan pengaruh fitur lainnya yang sangat minimal, menunjukkan ketergantungan model linear pada korelasi langsung berat badan. Sementara itu, Random Forest memperlihatkan distribusi kepentingan fitur yang lebih beragam; selain Weight (kg) yang tetap dominan, fitur BMI memberikan kontribusi yang cukup signifikan dibandingkan pada model lain. Pada model XGBoost, Weight (kg) menjadi faktor penentu utama yang diikuti oleh fitur aktivitas seperti Workout Frequency dan Physical Exercise sebagai pendukung akurasi. Serupa dengan XGBoost, LightGBM juga menunjukkan ketergantungan yang sangat tinggi pada fitur Weight (kg), di mana fitur pendukung lainnya memiliki pengaruh yang jauh lebih kecil terhadap stabilitas prediksi secara keseluruhan.

3.4. Modeling

Proses pelatihan dilakukan menggunakan 16.000 data latih yang telah melalui tahap normalisasi. Setiap algoritma dilatih secara efisien menggunakan subset 15 fitur terbaik terbaik hasil seleksi RFE agar model dapat fokus mempelajari pola hubungan yang paling relevan tanpa gangguan fitur redundan. Pendekatan ini memungkinkan model linear untuk mengandalkan hubungan linier antar fitur, sementara model ensemble dapat memaksimalkan kemampuannya dalam menangkap pola non-linear. Untuk mencapai performa yang paling stabil, dilakukan optimasi melalui GridSearchCV dengan 5-fold cross-validation menggunakan metrik R^2. Pada model Linear Regression, optimasi menunjukkan hasil terbaik dengan tetap menyertakan intercept. Untuk Random Forest, keseimbangan antara kedalaman pohon dan jumlah estimator dicapai pada n_estimators 100 dengan max_depth 20. Sementara itu, model XGBoost dan LightGBM bekerja paling optimal dengan laju pembelajaran menengah (learning rate 0.1) dan struktur pohon yang terkontrol untuk mencegah overfitting. Setelah parameter terbaik ditemukan, model disimpan sebagai best estimator untuk menghasilkan prediksi pada data uji. Analisis melalui Permutation Importance memperlihatkan bahwa seluruh model sangat bergantung pada fitur Weight (kg) sebagai prediktor utama. Meskipun demikian, terdapat perbedaan karakteristik di mana Random Forest memberikan bobot yang cukup besar pada fitur BMI, sedangkan XGBoost dan LightGBM lebih memanfaatkan fitur aktivitas fisik sebagai pendukung stabilitas prediksi.

3.5. Evaluation

Setelah keempat model telah dilatih, keempat model kemudian diuji menggunakan data uji (X_test_selected dan y_test). Prediksi nilai Calories dihasilkan melalui model.predict(X_test_selected),

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

kemudian dibandingkan dengan nilai aktual untuk menghitung tiga metrik evaluasi regresi, yaitu R^2 Score, Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE). Nilai metrik untuk masing-masing model disimpan pada results_eval dan diringkas dalam tabel result_df sehingga memudahkan perbandingan antar model. Hasil perbandingan performa evaluasi model dirangkum dalam Tabel 4.

Tabel 4. Hasil Evaluasi

Model	R^2	MAE	RMSE
Linear Regression	0.965026	80.945500	101.711033
Random Forest	0.961499	84.706590	106.716247
XGBoost	0.961151	84.932078	107.197916
LightGBM	0.964133	81.882680	103.001300

Pada Tabel 4 hasil evaluasi menunjukkan bahwa Linear Regression memberikan kinerja terbaik secara keseluruhan dengan $R^2 = 0.9650$, MAE = 80.95, dan RMSE = 101.71. Nilai R^2 yang mendekati 1 mengindikasikan bahwa model mampu menjelaskan sebagian besar variasi pada target Calories, sementara MAE dan RMSE yang paling kecil menunjukkan kesalahan prediksi rata-rata dan kesalahan kuadrat yang relatif rendah dibanding model lainnya. Pada kelompok model ensemble, LightGBM menjadi yang paling kompetitif dengan $R^2 = 0.9641$, MAE = 81.88, dan RMSE = 103.00, diikuti oleh Random Forest dengan $R^2 = 0.9615$, MAE = 84.71, RMSE = 106.72, serta XGBoost dengan $R^2 = 0.9612$, MAE = 84.93, dan RMSE = 107.20. Hasil evaluasi seluruh model menunjukkan performa tinggi dengan $R^2 > 0.96$ yang menandakan bahwa kombinasi fitur gaya hidup dan antropometrik dalam dataset cukup informatif untuk memprediksi kebutuhan kalori harian.

Kinerja model dievaluasi kembali setelah masing-masing algoritma memperoleh konfigurasi hiperparameter terbaik dari proses GridSearchCV (cv=3, scoring=R^2). Evaluasi dilakukan dengan cara menggunakan best estimator dari setiap model untuk memprediksi data uji yang telah disesuaikan dengan fitur terpilih hasil RFE masing-masing algoritma. Selain menghitung metrik R^2, MAE, dan RMSE, pada tahap ini juga dilakukan pengukuran waktu pelatihan (training time) untuk melihat efisiensi komputasi setiap model setelah optimasi. Hasil lengkap evaluasi ini disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi Setelah Hyperparameter Tuning

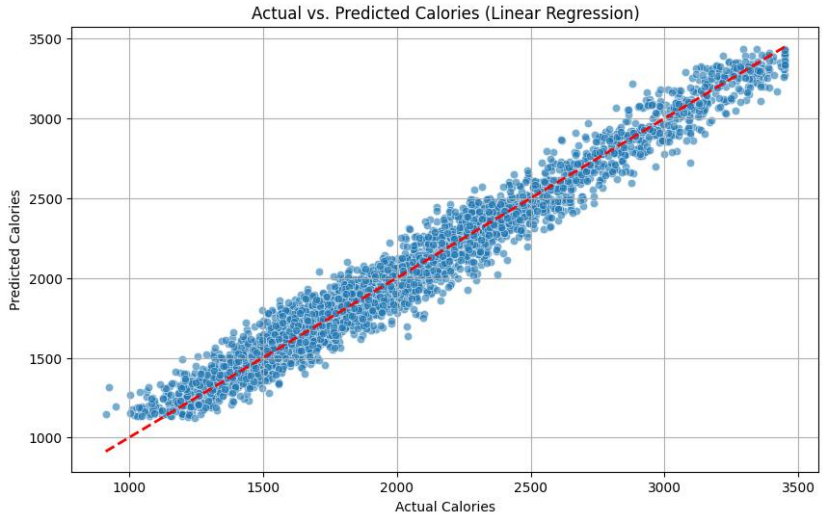
Model	R^2	MAE	RMSE	Training Time (s)
Linear Regression	0.965026	80.945500	101.711033	8.950357
Random Forest	0.962164	83.983755	105.791363	70.617932
XGBoost	0.964587	81.340287	102.347376	10.555561
LightGBM	0.964133	81.882680	103.001300	18.496122

Berdasarkan Tabel 5, proses hyperpatameter tuning berhasil meningkatkan akurasi seluruh model ensemble, di mana XGBoost mencatatkan peningkatan performa yang paling signifikan dengan nilai R^2 mencapai 0.9645 dan LightGBM stabil di angka 0.9641. Meskipun model berbasis pohon mengalami perbaikan, Linear Regression tetap mempertahankan posisinya sebagai model dengan kinerja terbaik secara keseluruhan, karena memiliki nilai R^2 tertinggi (0.9650) serta tingkat kesalahan prediksi terendah dengan MAE 80.94 dan RMSE

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

101.71. Dari sisi efisiensi komputasi, Linear Regression dan XGBoost terbukti paling praktis dengan waktu pelatihan masing-masing hanya 8.95 detik dan 10.55 detik, sementara Random Forest memerlukan waktu yang jauh lebih lama yaitu 70.61 detik meskipun tingkat akurasinya (R^2 0.9621) masih berada di bawah model lainnya.

Secara keseluruhan, Linear Regression tetap menjadi model yang paling unggul baik dari segi ketepatan prediksi maupun kecepatan waktu komputasi untuk dataset ini. Untuk memvisualisasikan seberapa dekat hasil prediksi model terbaik tersebut dengan nilai aktual di lapangan, disajikan grafik scatter plot pada Gambar 5.



Gambar 4. Scatter Plots (Actual vs. Predicted)

Gambar 4 menunjukkan plot sebaran antara nilai actual Calories dan nilai yang diprediksi oleh model Linear Regression. Terlihat bahwa titik-titik data tersebar sangat rapat mengikuti garis diagonal merah (garis referensi y=x), yang mengonfirmasi secara visual bahwa model memiliki tingkat akurasi yang sangat tinggi dan mampu memprediksi target dengan penyimpangan yang minimal.

4. DISCUSSIONS

Hasil tuning menggunakan GridSearchCV menunjukkan bahwa dampak optimasi hiperparameter bervariasi pada tiap model. Pada Linear Regression dan LightGBM, nilai metrik evaluasi tidak mengalami perubahan setelah tuning. Linear Regression tetap stabil dengan nilai R^2 0.965026, MAE 80.945500, dan RMSE 101.711033. Hal serupa terjadi pada LightGBM yang menetap pada R^2 0.964133, MAE 81.882680, dan RMSE 103.001300, mengindikasikan bahwa konfigurasi awal atau ruang parameter yang diuji sudah sangat mendekati optimal untuk dataset ini. Sebaliknya, perbaikan performa terlihat pada model ensemble lainnya. Random Forest mengalami sedikit peningkatan, di mana R^2 naik dari 0.961499 menjadi 0.962164, MAE turun dari 84.706590 menjadi 83.983755, dan RMSE membaik dari 106.716247 menjadi 105.791363. Peningkatan paling signifikan terjadi pada model XGBoost, dengan lompatan nilai R^2 dari 0.961151 menjadi 0.964587, serta penurunan tingkat error yang cukup besar pada MAE (menjadi 81.340287) dan RMSE (menjadi 102.347376). Hal ini menandakan bahwa proses optimasi sangat efektif dalam memaksimalkan kemampuan prediksi pada algoritma XGBoost.

Performa terbaik tetap dicapai oleh Linear Regression. Di antara model ensemble, hasil tuning menempatkan XGBoost sebagai yang paling kompetitif karena mampu mengungguli LightGBM baik pada metrik MAE maupun RMSE. Temuan bahwa model linear justru unggul mengindikasikan bahwa

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

setelah preprocessing dan seleksi fitur, hubungan antara fitur gaya hidup dan target Calories cenderung cukup linier dan stabil, sehingga model sederhana mampu memodelkan pola utama tanpa kompleksitas tinggi. Hal ini sejalan dengan studi terkait estimasi energi yang melaporkan bahwa pendekatan regresi linear dapat menghasilkan performa sangat tinggi dalam konteks tertentu ketika fitur yang digunakan relevan [19].

Jika dibandingkan dengan penelitian terdahulu yang sejenis, model Linear Regression yang diusulkan dalam penelitian ini mencapai tingkat akurasi yang lebih tinggi dengan nilai R^2 sebesar 0.9650, mengungguli studi terbaru oleh Tobin dkk. (2025) yang juga menggunakan pendekatan regresi linear berganda untuk memprediksi kepadatan energi dan mencapai R^2 maksimum sebesar 0.869. Meskipun beberapa studi melaporkan performa yang lebih tinggi pada boosting ketika menggunakan prosedur optimasi dan validasi yang ketat. Misalnya, XGBoost yang dioptimasi dengan Bayesian Optimization dan Nested Cross Validation dilaporkan mencapai R^2 yang sangat tinggi pada prediksi kalori berbasis aktivitas fisik [17], hasil penelitian ini tetap menunjukkan performa yang kuat dan konsisten ($R^2 > 0,96$ pada semua model), serta memperlihatkan bahwa tahapan praproses data yang terstruktur (penanganan outlier, encoding, normalisasi, dan RFE) mampu menghasilkan pemodelan yang stabil pada dataset gaya hidup.

Keunggulan Linear Regression tidak hanya terletak pada tingginya performa prediksi, melainkan juga pada trade-off yang ditawarkan. Karena sifatnya yang sederhana dan interpretable (mudah diinterpretasikan secara matematis), Linear Regression jauh lebih transparan dibandingkan model ensemble yang kompleks seperti XGBoost atau Random Forest, di mana proses pengambilan keputusannya sangat berlapis dan sulit diuraikan secara langsung. Karakteristik ini memberikan urgensi dan dampak yang tegas pada bidang Informatika, khususnya sebagai kontribusi terhadap sistem Computer Vision dan Medical Image Analysis. Dalam pengembangan sistem kesehatan digital masa depan, estimasi kalori akan diintegrasikan dengan pengenalan gambar makanan (image-based food recognition). Model regresi yang sangat efisien dan ringan ini menjadi pilihan paling ideal untuk ditanamkan pada aplikasi mobile health real-time, karena mampu memproses perhitungan kalori secara instan tepat setelah algoritma Computer Vision mendeteksi objek, tanpa membebani sumber daya memori maupun baterai perangkat genggam. Walaupun demikian, nilai performa yang sangat tinggi perlu dibahas secara hati-hati dalam konteks akademis karena bisa menimbulkan pertanyaan terkait potensi overfitting atau ketergantungan pada karakteristik dataset tertentu. Oleh karena itu, penelitian selanjutnya dapat memperkuat temuan dengan menguji kemampuan generalisasi model melalui validasi yang lebih robust, misalnya menggunakan k-fold cross-validation dan/atau validasi eksternal pada dataset lain atau pada kelompok populasi berbeda (berdasarkan usia, gender, maupun kategori BMI). Selain itu, studi lanjutan dapat melakukan ablation study untuk mengukur kontribusi masing-masing tahap dalam praproses data (penanganan outlier, standardisasi, dan seleksi fitur) terhadap peningkatan performa, sehingga diperoleh pemahaman yang lebih jelas mengenai komponen paling berpengaruh. Mengingat Linear Regression menunjukkan performa terbaik, pengembangan berikutnya juga dapat mengeksplorasi model linear ter-regularisasi seperti Ridge, Lasso, atau Elastic Net untuk meningkatkan stabilitas koefisien dan mengurangi potensi multikolinearitas tanpa mengorbankan interpretabilitas. Dari sisi evaluasi, analisis error yang lebih rinci per segmen (misalnya kelompok usia atau tingkat aktivitas) serta pemeriksaan residual dapat membantu mengidentifikasi kondisi ketika model cenderung mengalami over/under-estimation. Pengembangan ini diharapkan dapat meningkatkan keandalan model ketika diimplementasikan pada aplikasi pemantauan nutrisi dan gaya hidup berbasis data.

5. CONCLUSION

Penelitian ini menunjukkan bahwa kebutuhan kalori harian dapat diprediksi secara akurat menggunakan pendekatan pembelajaran mesin berbasis variabel gaya hidup ketika data diproses melalui

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

alur kerja yang terstruktur. Penerapan pembersihan data, transformasi fitur kategorikal, penanganan nilai ekstrem, standardisasi, serta seleksi fitur terbukti membantu membentuk representasi data yang lebih stabil, sehingga seluruh model yang diuji mampu mencapai performa tinggi. Dalam perbandingan model, Linear Regression memberikan hasil paling baik dan konsisten, mengindikasikan bahwa pola hubungan antara faktor antropometrik, pola makan, dan aktivitas fisik terhadap target kalori dapat ditangkap secara efektif oleh model yang lebih linier. Sementara itu, model ensemble khususnya boosting tetap kompetitif dan menunjukkan perbaikan setelah optimasi hiperparameter, menegaskan pentingnya penyetelan model untuk memperoleh konfigurasi terbaik.

Temuan ini memberikan kontribusi yang kuat pada bidang Informatika dengan menyajikan tolok ukur baru yang membuktikan bahwa model sederhana dan transparan mampu mencapai tingkat akurasi yang sangat tinggi dan kompetitif pada dataset terstruktur yang spesifik. Hal ini sekaligus menantang asumsi ketergantungan pada model kompleks yang memakan beban komputasi besar. Sebagai saran untuk penelitian lanjutan, pengembangan sistem diharapkan tidak hanya berhenti pada pengujian data tabular yang lebih beragam, melainkan dapat diintegrasikan dengan teknologi pengenalan citra (image recognition) untuk mendeteksi porsi dan jenis makanan. Integrasi pemodelan tabular gaya hidup dengan Computer Vision ini akan menciptakan sistem prediksi hybrid multi-modal yang sangat efisien dan ideal untuk diterapkan pada aplikasi mobile health secara real-time.

REFERENCES

[1] N. Afni and Z. Al Faiqoh, "Perbedaan Asupan Energi Makronutrien, Aktivitas Fisik Dan Status Gizi Pada Siswa Di Sma Wahid Hasyim Model Lamongan Yang Bermukim Di Pondok Pesantren Dan Yang Bermukim Di Rumah," *Heal. Tadulako J. (Jurnal Kesehat. Tadulako)*, vol. 10, no. 2, pp. 306–315, 2024.

[2] M. S. N. Lugia Wanda, "3889-11021-1-Pb," *Hub. Akt. Fis. Energi,Dansarapanpagi Dengan Kejadianoverweightpada Siswa Sma*, vol. 17, no. 2, pp. 1–9, 2021.

[3] M. G. Pantaleon, Y. Petrika, A. U. Zogara, Desi, dan M. Niron, "Hubungan asupan energi dan zat gizi serta pengetahuan dengan status gizi pada remaja di Kota Kupang," *SAGO: Gizi dan Kesehatan*, vol. 6, no. 2, pp. 301–308, 2025, doi: 10.30867/gikes.v6i2.2388.

[4] D. M. Sari, "Asupan Energi, Kebiasaan Olahraga dan Status Gizi pada Remaja di Inderalaya," *Jurnal Gizi Dietetik*, vol. 4, no. 2, pp. 151–157, 2025, doi: 10.25182/jigd.2025.4.2.151-157.

[5] M. Irwanda, D. Suryani, A. Krisnasary, dan Yandrizal, "Gambaran Asupan Energi, Zat Gizi Makro Dan Status Gizi Remaja Di SMP N 14 Kota Bengkulu Tahun 2022," *AKSARA: Jurnal Ilmu Pendidikan Nonformal*, vol. 9, no. 1, pp. 199–208, 2023, doi: 10.37905/aksara.9.1.199-208.2023.

[6] D. Mukhtar, K. A. D. A. Ridwan, and H. L. Fitriani, "Impact of Calorie Intake on Cardiovascular Disease Risk Factors for Young Adults Working from Home During the COVID-19 Pandemic," vol. 7, no. 1, pp. 47–55, 2025, doi: 10.35790/msj.v7i1.52343.

[7] Y. W. A. Rustam and Hendra Gunawan, "Perancangan Aplikasi Perhitungan Kebutuhan Kalori Tubuh Harian Berdasarkan Asupan Konsumsi Makanan Menggunakan Logika Fuzzy," *Inf. (Jurnal Inform. dan Sist. Informasi)*, vol. 14, no. 2, pp. 94–109, 2022, doi: 10.37424/informasi.v14i2.174.

[8] R. Riswanto, A. Ahmad, H. Hazriani, and D. Tribuana, "Deteksi Kalori Makanan Tradisional Indonesia Menggunakan Metode Single Shot Multibox Detector (SSD)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 819–829, 2024, doi: 10.57152/malcom.v4i3.1332.

[9] S. A. Oktavianti, Y. Divayana, and I. G. A. P. Raka Agung, "Aplikasi NutriNeeds dalam penentuan kebutuhan kalori harian bagi penderita diabetes melitus," *Jurnal SPEKTRUM*, vol. 8, no. 2, pp. 48–54, Jul. 2021, doi: 10.24843/SPEKTRUM.2021.v08.i02.p6.

[10] A. Z. Ulhaq, A. Z. Adilukito, S. M. P. G. Neru, and M. D. Agisfio, "Aplikasi Pencatatan Kalori Harian Berbasis Android Dengan Arsitektur MVVM," *Computer Science (CO-SCIENCE)*, vol. 5, no. 1, pp. 26–34, 2025, doi: 10.31294/coscience.v5i1.3443.

[11] J. Zheng, J. Wang, J. Shen, and R. An, "Artificial Intelligence Applications to Measure Food and

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

Nutrient Intakes: Scoping Review," *J. Med. Internet Res.*, vol. 26, p. e54557, 2024, doi: 10.2196/54557.

[12] S. K. Aydın, R. H. Ali, S. Faiz, and T. A. Khan, "An Integrated AI Framework for Personalized Nutrition Using Machine Learning and Natural Language Processing for Dietary Recommendations," *Applied Sciences*, vol. 15, no. 17, Art. no. 9283, 2025, doi: 10.3390/app15179283.

[13] C. Y. S. Ang, M. B. M. Nor, N. S. Nordin, T. Z. Kyi, A. Razali, and Y. S. Chiew, "Methods for estimating resting energy expenditure in intensive care patients: A comparative study of predictive equations with machine learning and deep learning approaches," *Computer Methods and Programs in Biomedicine*, vol. 262, p. 108657, Apr. 2025, doi: 10.1016/j.cmpb.2025.108657.

[14] R. Ruede, V. Heusser, L. Frank, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "Multi-task learning for calorie prediction on a novel large-scale recipe dataset enriched with nutritional information," *Proc. - Int. Conf. Pattern Recognit.*, pp. 4001–4008, 2021, doi: 10.1109/ICPR48806.2021.9412839.

[15] S. Mujiyono, U. P. Sanjaya, I. S. Wibisono, and H. Setyowati, "Prediksi Fluktuasi Berat Badan Berdasarkan Pola Hidup Menggunakan Model XGBoost dan Deep Learning," *J. Algoritm.*, vol. 22, no. 1, pp. 221–233, 2025, doi: 10.33364/algoritma/v.22-1.2253.

[16] S. Wibawa, A. Suherman, K. Sultoni, Jajat, Y. Ruhayati, and W. D. Nuryanti, "Estimasi Kalori Expenditure Berdasarkan Accelerometer ActiGraph dan Ergocycle," *Jambura Journal of Sports Coaching*, vol. 7, no. 1, pp. 118–124, Jan. 2025.

[17] B. Budiman, N. Alamsyah, T. Parama Yoga, R. Y. Rakhman Alamsyah, and E. Setiana, "XGBoost optimization using hybrid Bayesian optimization and nested cross validation for calorie prediction," *TELKOMNIKA (Telecommunication Comput. Electron. Control)*., vol. 23, no. 3, p. 694, 2025, doi: 10.12928/TELKOMNIKA.v23i3.26554.

[18] C. E. Sukmawati, A. Fitri, N. Masruriyah, and A. R. Juwita, "Efektivitas algoritma AdaBoost dan XGBoost pada dataset obesitas populasi dewasa," *Jambura J. Informatics*, vol. 6, no. 2, pp. 101–111, 2024. doi: 10.37905/jji. v6i2.25194.

[19] T. A. Adjuik, N. A. A. Boi-Dsane, and B. A. Kehinde, "Enhancing dietary analysis: Using machine learning for food caloric and health risk assessment," *J. Food Sci.*, vol. 89, no. 11, pp. 8006–8021, 2024, doi: 10.1111/1750-3841.17421.

[20] G. Tobin, A. Schuhmacher, T. Górecki, and Ł. Smaga, "The development and evaluation of multiple regression equations based on four common nutritional analysis packages to predict the metabolisable energy density of diets fed to grower / finisher and adult pigs and their use for rat and mouse diets," *Br. J. Nutr.*, vol. 133, no. 4, pp. 433–455, 2025, doi: 10.1017/S0007114525000042.

[21] P. PAULRAJ, P. M. P.MANOJ, and S. S. S.SELVABHARATHI, "Calorie Intake Prediction Using Machine Learning for Personalized Food Recommendations," *Int. J. Creat. Res. Thoughts*, vol. 13, no. 7, pp. 687–693, 2025.

[22] S. S. Ratnakar and S. Vidya, "Calorie Burn Predection using Machine Learning," *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, vol. 9, no. 6, pp. 781–787, Jun. 2022, doi: 10.17148/IARJSET.2022.96125

[23] S. Devi K. A., G. S, M. Basavaraj, and M. V. S., "Hybrid Model Analysis for Calorie Prediction Using Ensemble Learning Techniques: XGBoost and Random Forest," *International Journal of Environmental Sciences*, vol. 11, no. 6, pp. 3024–3031, 2025, doi: 10.64252/1zx66k89.

[24] N. K. Hamzidah, A. Ulandari, M. M. Parenreng, and N. Ichzan As, "Evaluasi Kinerja Aplikasi Mobile Penghitung Kalori Makanan Berbasis Algoritma CNN-YOLO (Performance Evaluation of Food Calorie Counter Mobile Application Based on CNN-YOLO Algorithms)," *Jambura Journal of Electrical and Electronics Engineering*, vol. 7, no. 2, pp. 253–263, 2025, doi: 10.37905/jjeee.v7i2.30595.

[25] P. Yarde, D. Bordoloi, R. M. Chavan, V. Vekariya, H. Patil, and L. Natrayan, "A Deep Learning Neural Network-based System for Food Recognition and Calorie Estimation," *2023 3rd Int. Conf. Innov. Mech. Ind. Appl.*, no. Icimia, pp. 1551–1558, 2023, doi: 10.1109/ICIMIA60377.2023.10426548.

Vol. 7, No. 3, June 2026, Page. 2841-2856
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5621

[26] M. Ogishi, H. Tanabe, and K. Yanai, "Diffusion-Guided 3D-Aware Calorie Estimation from a Single Food Image," in *Proc. 1st International Workshop on Multi-modal Food Computing (MMFood '25)* (co-located with *ACM Multimedia 2025*), 2025, doi: 10.1145/3746264.3760487.

[27] E. S. Sintiya, S. R. Amanda, C. Bella Vista, and A. Nugroho Pramudhita, "Implementasi Machine Learning dalam Sistem Prediksi dan Rekomendasi Program Diet Terintegrasi LLM," *Jurnal Nasional Teknologi dan Sistem Informasi (TEKNOSI)*, vol. 11, no. 2, pp. 144–151, Sep. 2025, doi: 10.25077/TEKNOSI.v11i2.2025.144-151.

[28] U. Nadifa, R. Deddy, R. Dako, A. I. Tolago, and R. Hidayat, "Efektivitas Optimasi Hyperparameter Dalam Prediksi Pembakaran Kalori : Data Aktivitas Fisik," vol. 7, pp. 1–8, 2025, doi: 10.32528/elkom.v7i2.22636191.

[29] N. Fosua, C. Courtney, O. Toole, and A. Jalali, "International Journal of Medical Informatics Comparing logistic regression and machine learning for obesity risk prediction : A systematic review and meta-analysis," *Int. J. Med. Inform.*, vol. 199, p. 105887, Jul. 2025, doi: 10.1016/j.ijmedinf.2025.105887.

[30] Y. Lu, C. Chen, J. Qiu, Q. Ji, L. Zhou, and H. Xiong, "Systematic review and comparison of machine learning and conventional statistical models for predicting cardiovascular events in dialysis patients," *Ren. Fail.*, vol. 47, no. 1, p. 2587490, Dec. 2025, doi: 10.1080/0886022X.2025.2587490.

[31] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh tahapan preprocessing terhadap model IndoBERT dan IndoBERTweet untuk mendeteksi emosi pada komentar akun berita Instagram," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.

[32] A. Fitrianto, A. Kholifatunnisa, and A. Kurnia, "Comparing Outlier Detection Methods : An Application on Indonesian Air Quality Data," vol. 9, no. 2, pp. 341–351, 2024, doi: 10.18860/ca.v9i2.29434.

[33] A. M. Priyatno, T. Widiyaningtyas, I. Engineering, and U. N. Malang, "A Systematic Literature Review: Recursive Feature Elimination Algorithms," vol. 9, no. 2, pp. 196–207, 2024, doi: 10.33480/jitk.v9i2.5015.

[34] O. Bulut, B. Tan, and E. Mazzullo, "Benchmarking Variants of Recursive Feature Elimination : Insights from Predictive Tasks in Education and Healthcare," pp. 1–21, 2025, doi: 10.3390/info16060476.

[35] W. Nugraha and A. Sasongko, "Hyperparameter Tuning on Classification Algorithm with Grid Search," *SISTEMASI: Jurnal Sistem Informasi*, vol. 11, no. 2, pp. 391–401, May 2022, doi: 10.32520/stmsi.v11i2.1750.

[36] D. Chicco, M. J. Warrens, and G. Jurman, "The coef fi cient of determination R-squared is more informative than SMAPE , MAE , MAPE , MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[37] F. Yasin, M. Raafi, D. Aldo, and M. A. Amrustian, "Multivariate Forecasting of Paddy Production : A Comparative Study of Machine Learning Models," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 3, pp. 1431–1442, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4681.

[38] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (MAE) into systematic and unsystematic components," *PLoS ONE*, vol. 18, no. 2, p. e0279774, 2023, doi: 10.1371/journal.pone.0279774.