Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Evaluasi Komparatif Metode Feature Selection pada XGBoost Regression untuk Prediksi Panjang Siklus Menstruasi

Shofiatuz Zulfia*, Mula Agung Barata, Ifnu Wisma Dwi Prastya
Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri, Bojonegoro, Indonesia
Email: 1,*shofiatuzzulfia@gmail.com, 2mula.ab26@gmail.com, 3ifnuprastya@unugiri.ac.id
Email Penulis Korespondensi: shofiatuzzulfia@gmail.com*
Submitted: **21/01/2026**; Accepted: **24/02/2026**; Published: **31/03/2026**

Abstrak−Panjang siklus menstruasi menjadi indikator utama dalam kesehatan reproduksi perempuan, namun perbedaan karakteristik individu dan ketidakteraturan siklus menyulitkan proses prediksi secara manual. Kondisi tersebut mendorong perlunya pendekatan berbasis data yang mampu menghasilkan prediksi panjang siklus menstruasi secara akurat dan konsisten. Penelitian ini bertujuan untuk melakukan evaluasi komparatif berbagai metode *feature selection* pada algoritma *XGBoost Regression* dalam memprediksi panjang siklus menstruasi. Dataset penelitian diperoleh dari Kaggle dan terdiri atas 162 data yang mencakup atribut fisiologis dan demografis perempuan. Tahapan penelitian meliputi *preprocessing* data, normalisasi menggunakan *StandardScaler*, pembagian data latih dan data uji dengan rasio 80:20, serta validasi 10-fold *cross-validation* untuk menguji stabilitas model. Empat skenario pemodelan dievaluasi, yaitu tanpa *feature selection* sebagai *baseline*, *forward selection*, *backward elimination*, dan *optimized selection* berbasis *ensemble feature selection* dari lima metode seleksi fitur. Hasil evaluasi menunjukkan bahwa metode *forward selection* memberikan performa terbaik dengan nilai R^2 sebesar 0,9005, RMSE 1,45 hari, MAE 0,57 hari, dan MAPE 1,73% (kesalahan relatif rata-rata < 2% terhadap panjang siklus 25-30 hari), serta meningkatkan nilai R^2 sebesar 0,1696 poin (dari 0,7309 menjadi 0,9005), setara dengan peningkatan relatif 23,2% terhadap nilai baseline. Temuan ini menunjukkan bahwa pemilihan metode *feature selection* yang tepat berpengaruh terhadap peningkatan performa prediktif dan stabilitas model *XGBoost Regression* dalam prediksi panjang siklus menstruasi.

Kata Kunci: *Feature selection*; *Forward* selection; Kesehatan reproduksi; Prediksi siklus menstruasi; *XGBoost Regression*.

Abstract−Menstrual cycle length is a key indicator of women's reproductive health, but differences in individual characteristics and cycle irregularities complicate manual prediction. This situation necessitates a data-driven approach capable of accurately and consistently predicting menstrual cycle length. This study aims to conduct a comparative evaluation of various feature selection methods in the XGBoost Regression algorithm in predicting menstrual cycle length. The research dataset was obtained from Kaggle and consists of 162 data points covering women's physiological and demographic attributes. The research stages include data preprocessing, normalization using StandardScaler, dividing the training data and test data with a ratio of 80:20, and 10-fold cross-validation to test model stability. Four modeling scenarios were evaluated, namely without feature selection as a baseline, forward selection, backward elimination, and optimized selection based on ensemble feature selection from five feature selection methods. The evaluation results showed that the forward selection method provided the best performance with an R^2 value of 0.9005, RMSE of 1.45 days, MAE of 0.57 days, and MAPE of 1.73% (average relative error <2% against a cycle length of 25-30 days), and increased the R^2 value by 0.1696 points (from 0.7309 to 0.9005), equivalent to a relative increase of 23.2% over the baseline value. These findings indicate that choosing the right feature selection method has an impact on improving the predictive performance and stability of the XGBoost Regression model in predicting menstrual cycle length.

Keywords: Feature selection; Forward selection; Menstrual cycle prediction; Reproductive health; XGBoost regression.

1. PENDAHULUAN

Kesehatan reproduksi perempuan merupakan salah satu aspek fundamental dalam kajian kesehatan masyarakat dan kedokteran preventif karena berhubungan langsung dengan kualitas hidup dan kesejahteraan perempuan sepanjang siklus kehidupannya [1], [2]. Salah satu indikator utama yang sering digunakan untuk menilai kondisi kesehatan reproduksi perempuan adalah panjang siklus menstruasi, yang didefinisikan sebagai jumlah hari dari hari pertama menstruasi hingga hari pertama menstruasi berikutnya dalam satu siklus [3]. Parameter ini mencerminkan keseimbangan hormonal dan kondisi fisiologis tubuh, sehingga ketidakteraturannya dikaitkan dengan gangguan hormonal, stres tinggi, aktivitas fisik tidak seimbang, serta risiko gangguan reproduksi [4]. Oleh karena itu, pemantauan dan prediksi panjang siklus menstruasi secara akurat memiliki peran strategis dalam mendukung deteksi dini gangguan kesehatan reproduksi, perencanaan fertilitas, serta pengembangan layanan kesehatan berbasis teknologi digital [5].

Namun demikian, panjang siklus menstruasi memiliki karakteristik yang sangat bervariasi antarindividu dan dipengaruhi oleh faktor fisiologis maupun lingkungan [6]. Variabilitas ini menyebabkan pendekatan prediksi berbasis perhitungan statistik sederhana memiliki keterbatasan dalam menangkap pola kompleks dan nonlinear [7]. Hal ini mendorong perlunya pendekatan berbasis data yang mampu memodelkan pola secara lebih adaptif dan sistematis [8], [9]. Dalam beberapa tahun terakhir, pendekatan *machine learning* berkembang pesat dalam kesehatan reproduksi karena kemampuannya menangani data multidimensi dan pola nonlinear [10], [11].

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Sejumlah penelitian menunjukkan bahwa *machine learning* memberikan performa lebih baik dibandingkan metode statistik klasik dalam analisis dan prediksi siklus menstruasi [8], [9], [10]. Rego [6] mengembangkan model deret waktu berbasis data historis untuk memprediksi panjang siklus menstruasi dan menunjukkan kemampuannya menangkap pola siklus secara adaptif. Sementara itu, Khairunisa dkk. [8], [12] melakukan perbandingan beberapa algoritma *machine learning* dalam prediksi siklus menstruasi dan melaporkan bahwa model berbasis *tree* dan *ensemble* memberikan performa yang lebih stabil pada data menstruasi yang heterogen. Penelitian lain juga mengonfirmasi efektivitas penerapan *machine learning* dalam aplikasi pelacakan periode, kesuburan, dan ovulasi [9], [10].

Di antara berbagai algoritma *ensemble learning*, *XGBoost Regression* (*eXtreme Gradient Boosting*) dikenal luas karena efisiensi komputasi, kemampuan regularisasi, serta performanya yang konsisten pada berbagai domain prediktif [13], [14]. Algoritma ini telah banyak digunakan dalam berbagai studi prediksi, termasuk prediksi kualitas udara [15], diagnosis penyakit [16], serta aplikasi prediktif berbasis data kesehatan lainnya [17]. Keunggulan utama *XGBoost* terletak pada kemampuannya memodelkan hubungan nonlinear yang kompleks serta mengendalikan *overfitting* melalui regularisasi yang terintegrasi dalam proses pelatihan model [13], [17].

Meskipun *XGBoost Regression* memiliki performa yang kuat, efektivitas model ini sangat dipengaruhi oleh kualitas dan relevansi fitur yang digunakan [18]. Beberapa penelitian melaporkan bahwa penggunaan seluruh fitur tanpa proses seleksi dapat meningkatkan kompleksitas model, memperbesar risiko *overfitting*, serta menurunkan kemampuan generalisasi, terutama pada dataset berukuran kecil hingga menengah [19], [20]. Hal ini menunjukkan bahwa keberhasilan pemodelan tidak hanya ditentukan oleh pemilihan algoritma, tetapi juga oleh strategi optimasi fitur yang diterapkan sebelum proses pelatihan model [21].

Dalam konteks tersebut, *feature selection* menjadi tahap penting dalam proses pemodelan *machine learning* karena berfungsi untuk meningkatkan efisiensi dan kinerja model [18]. Metode *feature selection* bertujuan untuk mengidentifikasi subset fitur yang paling relevan dan informatif, sehingga mampu meningkatkan performa prediksi, stabilitas, serta interpretabilitas model [22]. Berbagai pendekatan *feature selection* telah dikembangkan, di antaranya *Forward Selection* yang bekerja dengan menambahkan fitur secara bertahap berdasarkan kontribusi peningkatan performa model, sedangkan *Backward Elimination* menghilangkan fitur yang dianggap kurang relevan dari keseluruhan fitur yang tersedia [23]. *Forward Selection* dan *Backward Elimination* termasuk dalam kategori *wrapper methods* karena evaluasi fitur dilakukan berdasarkan performa model secara iteratif [24].

Penelitian ini juga menerapkan *Optimized Feature Selection*, berbasis *ensemble feature selection* dari lima metode seleksi fitur, yang menggabungkan *korelasi Pearson* untuk mengukur hubungan linear, *mutual information* untuk mendeteksi ketergantungan nonlinear, *F-test (ANOVA)* untuk mengevaluasi relevansi statistik, *feature importance* berbasis *XGBoost* untuk menilai kontribusi relatif fitur, dan *Recursive Feature Addition (RFA)* untuk evaluasi performa iteratif [25]. Pendekatan ini bertujuan memperoleh subset fitur yang relevan secara statistik sekaligus optimal bagi performa regresi [26].

Meskipun berbagai studi telah mengkaji penerapan *machine learning*, *XGBoost Regression*, dan *feature selection* pada beragam domain, penelitian yang secara khusus melakukan evaluasi komparatif antara model *XGBoost Regression* tanpa seleksi fitur (*baseline*) dan berbagai metode *feature selection* dalam konteks prediksi panjang siklus menstruasi masih relatif terbatas [6], [8]. Sebagian besar penelitian sebelumnya berfokus pada perbandingan algoritma atau penerapan satu metode seleksi fitur tanpa analisis komparatif yang sistematis terhadap dampaknya pada performa model prediksi siklus menstruasi [12].

Berdasarkan permasalahan dan celah penelitian tersebut, penelitian ini bertujuan untuk melakukan evaluasi komparatif berbagai metode *feature selection* pada algoritma *XGBoost Regression* dalam memprediksi panjang siklus menstruasi [13], [18]. Metode yang dievaluasi meliputi model *baseline* tanpa *feature selection*, *Forward Selection*, *Backward Elimination*, serta *Optimized Selection* [23], [25]. Dataset yang digunakan adalah *Menstrual Cycle Dataset* dari Kaggle yang terdiri dari 162 sampel [27]. Ukuran dataset yang relatif kecil menjadi salah satu keterbatasan penelitian ini, sehingga hasil evaluasi perlu diinterpretasikan dengan mempertimbangkan potensi variabilitas sampel. Diharapkan penelitian ini dapat memberikan kontribusi dalam bentuk pemahaman empiris mengenai pengaruh pemilihan metode *feature selection* terhadap performa model berdasarkan metrik regresi dan stabilitas model *XGBoost Regression*, serta menjadi dasar pengembangan sistem prediksi siklus menstruasi yang lebih andal dalam mendukung pemantauan kesehatan reproduksi perempuan [4], [5].

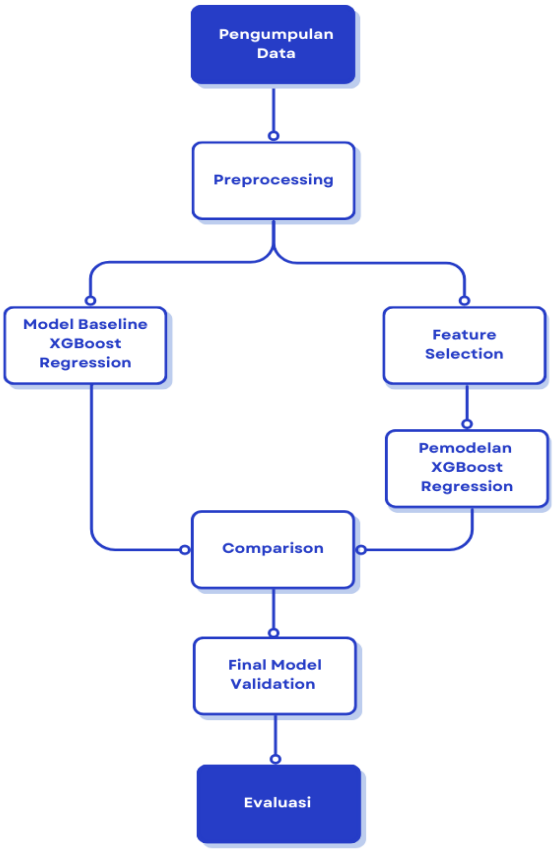
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian dirancang secara sistematis untuk mengevaluasi pengaruh berbagai metode *feature selection* terhadap kinerja *XGBoost Regression* dalam memprediksi panjang siklus menstruasi. Penelitian diawali

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

dengan pengumpulan data dan *preprocessing* untuk memastikan kualitas data yang digunakan. Selanjutnya, dilakukan pembangunan model *baseline XGBoost Regression* tanpa seleksi fitur sebagai acuan performa awal. Setelah itu, beberapa metode *feature selection* diterapkan untuk menghasilkan *subset* fitur yang berbeda, yang kemudian digunakan dalam pemodelan *XGBoost Regression*. Kinerja seluruh konfigurasi model dievaluasi menggunakan metrik regresi dan divalidasi melalui *cross-validation*. Hasil evaluasi tersebut dibandingkan untuk menentukan model terbaik, yang selanjutnya diuji kembali pada tahap *final model validation* guna menilai kemampuan generalisasi. Alur penelitian ini disajikan secara visual pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle berjudul *"Menstrual Cycle Dataset"*, yang berisi data fisiologis dan demografis perempuan terkait siklus menstruasi [27]. Variabel target dalam penelitian ini adalah Length_of_cycle, yang didefinisikan sebagai jumlah hari dari hari pertama menstruasi hingga hari pertama menstruasi berikutnya dalam satu siklus, dengan rentang 22–40 hari dan rata-rata 29,28 ± 4,19 hari. Dataset ini terdiri dari 162 sampel observasi dan mencakup 13 kolom fitur, di antaranya:

Tabel 1. Fitur Dataset

No	Fitur	Deskripsi
1	number_of_peaks	Jumlah puncak perdarahan menstruasi dalam satu siklus
2	Age	Usia responden (tahun)
3	Length_of_cycle	Panjang total siklus menstruasi (hari)
4	Estimated_day_of_ovulation	Hari perkiraan ovulasi sejak awal menstruasi.
5	Length_of_Leutal_Phase	Lama fase luteal (hari), dari ovulasi hingga menstruasi berikutnya.
6	Length_of_menses	Lama perdarahan menstruasi (hari).
7	Unusual_Bleeding	Perdarahan tidak normal di luar menstruasi (Ya/Tidak).
8	Height	Tinggi badan (dikonversi ke cm).
9	Weight	Berat badan (kg).
10	Income	Pendapatan bulanan (INR)
11	BMI	Indeks Massa Tubuh (kg/m²).
12	Mean_of_length_of_cycle	Rata-rata panjang siklus selama satu tahun (hari).
13	Menses_score	Skor nyeri menstruasi (1–5; 1 = tidak nyeri, 5 = sangat nyeri).

Sampel data dalam penelitian ini disajikan pada Tabel 2, yang menampilkan 10 observasi awal dan observasi ke-162 sebagai representasi akhir dataset.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Tabel 2. Sampel Dataset

No	number_of _peaks	Age	Length_of _cycle	Estimated_day_of _ovulation	Length_of_Leutal _Phase	...	Menses_ score
1	3	18	27	14	9	…	4
2	4	18	25	17	10	…	5
3	2	19	30	17	13	…	2
4	3	19	28	16	14	…	3
5	2	19	35	18	15	…	5
6	5	19	30	15	10	…	2
7	3	19	40	10	12	…	3
8	2	20	30	14	10	…	5
9	3	18	30	16	14	…	5
10	3	19	27	15	16	…	3
...	...	...	...	...	...	…	...
162	3	19	28	7	14	…	5

2.3 *Preprocessing* Data

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum pemodelan. Proses ini mencakup pengecekan tipe data, duplikasi, *missing value*, dan *outlier*. Dataset terdiri dari 162 observasi dengan 13 variabel, meliputi 8 variabel numerik bertipe integer, 3 variabel numerik bertipe *float*, dan 2 variabel kategorikal bertipe *object*. Pemeriksaan duplikasi dilakukan untuk mengidentifikasi observasi yang identik. Duplikasi yang terdeteksi dihapus untuk menghindari bias dalam pelatihan model. Penanganan *missing value* dilakukan secara sistematis untuk menjaga kualitas dan konsistensi data. Pada variabel numerik, nilai yang hilang diimputasi menggunakan median untuk mengurangi pengaruh outlier. Untuk variabel kategorikal, imputasi dilakukan menggunakan modus, yaitu kategori yang paling sering muncul. Sementara itu, apabila suatu variabel memiliki proporsi *missing value* lebih dari 30%, variabel tersebut dihapus dari dataset karena dianggap tidak memiliki informasi yang cukup dan berpotensi menurunkan kualitas model.

Selanjutnya, identifikasi *outlier* dilakukan menggunakan metode *Interquartile Range* (IQR) dengan batas bawah $Q1 - 1.5 \times IQR$ dan batas atas $Q3 + 1.5 \times IQR$. *Outlier* dievaluasi berdasarkan batas fisiologis dan tetap dipertahankan jika masih wajar secara medis. Kemudian transformasi fitur diterapkan melalui konversi fitur kategorikal ke bentuk numerik menggunakan *label encoding* untuk variabel biner (*Unusual_Bleeding*: 'no' = 0, 'yes' = 1) serta konversi satuan *Height* dari format teks (misalnya "5'3"") ke sentimeter (cm) menggunakan faktor konversi standar 1 kaki = 30,48 cm dan 1 inci = 2,54 cm.

Setelah itu dataset dibagi menjadi dua bagian utama, yaitu data latih (*training set*) dan data uji (*testing set*), dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Pembagian dilakukan secara acak (*random split*) menggunakan fungsi train_test_split dari pustaka *scikit-learn* dengan parameter *test_size*=0.2 dan *random_state* yang ditetapkan untuk memastikan reproduktibilitas eksperimen.

Setelah pembagian, seluruh fitur numerik dinormalisasi menggunakan *StandardScaler* untuk menghasilkan distribusi dengan rata-rata nol dan simpangan baku satu. Transformasi yang digunakan adalah:

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

di mana X adalah nilai fitur asli, μ adalah rata-rata fitur pada data latih, dan σ adalah simpangan baku fitur pada data latih. Meskipun XGBoost secara teoretis tidak memerlukan normalisasi karena berbasis pohon keputusan [28]. Langkah ini diambil terutama untuk konsistensi dalam visualisasi, kompatibilitas dengan metode seleksi fitur berbasis jarak (seperti *mutual information* dan F-test), serta memudahkan interpretasi koefisien dalam analisis komparatif [22].

2.4 Pemodelan Baseline *XGBoost Regression*

Model *baseline XGBoost Regression* dibangun menggunakan seluruh fitur numerik hasil preprocessing tanpa penerapan metode seleksi fitur. Dataset dibagi secara acak menjadi 80% data latih dan 20% data uji. Pelatihan model dilakukan pada data latih, sedangkan evaluasi performa dilakukan pada data uji untuk mengukur kemampuan generalisasi terhadap data yang tidak terlihat selama proses pelatihan. Selain itu, model divalidasi menggunakan skema 10-fold *cross-validation* pada data latih dengan metrik RMSE untuk menilai stabilitas prediksi. Nilai rata-rata RMSE merepresentasikan tingkat kesalahan umum model, sedangkan standar deviasi menunjukkan konsistensi error antar fold. Model ini berfungsi sebagai acuan performa awal untuk mengevaluasi penerapan *feature selection* sehingga mampu meningkatkan performa prediksi panjang siklus menstruasi.

2.5 Metode *Feature Selection*

Penelitian ini menerapkan tiga pendekatan seleksi fitur, yaitu Forward Selection, Backward Elimination, dan Optimized Selection, yang masing-masing dikombinasikan dengan XGBoost Regression untuk membentuk konfigurasi model yang berbeda.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

2.5.1 *Forward Selection + XGBoost Regression*

Forward Selection diterapkan sebagai metode *wrapper* yang bekerja secara bertahap dengan membangun model mulai dari subset fitur kosong. Pada setiap iterasi, satu fitur dari kumpulan fitur yang tersisa ditambahkan ke dalam model berdasarkan kontribusinya dalam meningkatkan nilai koefisien determinasi (R^2) pada data validasi. Proses penambahan fitur ini dilakukan secara berulang hingga tidak lagi menghasilkan peningkatan performa yang substansial, yang ditetapkan sebagai kriteria *early stopping*. Proses ini dijalankan hingga maksimal 10 fitur atau hingga tidak ada peningkatan R^2. Dalam eksperimen ini, proses berhenti pada 5 fitur karena tidak ada peningkatan substansial setelah iterasi ke-5.

2.5.2 *Backward Elimination + XGBoost Regression*

Backward Elimination diimplementasikan sebagai metode seleksi fitur berbasis penghapusan bertahap yang diawali dengan penggunaan seluruh 11 fitur yang tersedia. Pada setiap iterasi, satu fitur dengan nilai *feature importance* terendah dihilangkan dari subset fitur. Setelah penghapusan, kinerja model dievaluasi menggunakan validasi silang dengan memperhatikan perubahan nilai koefisien determinasi (R^2). Jika penurunan R^2 yang terjadi tidak melebihi batas toleransi sebesar 5%, maka fitur tersebut secara permanen dihapus dan proses dilanjutkan ke iterasi berikutnya. Sebaliknya, apabila penurunan performa melebihi ambang batas yang ditetapkan, proses eliminasi dihentikan. Dalam eksperimen ini, proses berhenti setelah 7 fitur dihapus, menyisakan 4 fitur yang paling informatif berdasarkan stabilitas prediksi.

2.5.3 *Optimized Selection + XGBoost Regression*

Pendekatan ini menggunakan *ensemble feature selection* dengan menggabungkan lima metode, yaitu korelasi Pearson, mutual information, F-test (ANOVA), *feature importance* XGBoost, dan *Recursive Feature Addition* (RFA). Setiap metode menghasilkan skor untuk setiap fitur, yang kemudian dinormalisasi ke [0,1] dan dirata-ratakan menjadi *ensemble score*. Selanjutnya, dilakukan pencarian subset optimal dengan menguji kombinasi fitur berdasarkan peringkat *ensemble score*, menggunakan 5-Fold Cross-Validation untuk mengevaluasi R^2 pada setiap ukuran subset (dari 3 hingga 11 fitur). Subset dengan CV R^2 tertinggi dipilih sebagai fitur optimal. Dalam eksperimen ini, 6 fitur terpilih sebagai konfigurasi terbaik.

2.6 Pemodelan *XGBoost Regression*

Model utama yang digunakan dalam penelitian ini adalah XGBoost Regression, yaitu algoritma *gradient boosting* yang membangun model secara aditif dari kumpulan pohon keputusan lemah [17]. Fungsi prediksi model XGBoost dirumuskan sebagai:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (2)$$

di mana $\hat{y}_i$ merupakan prediksi untuk sampel ke-i, Kadalah jumlah pohon keputusan, $\mathcal{F}$ adalah himpunan fungsi pohon (CART), dan f_k menyatakan fungsi pohon ke-k.
Fungsi objektif yang diminimalkan selama pelatihan model didefinisikan sebagai:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

Fungsi *loss* regresi berupa *squared error* dan fungsi regularisasi:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \quad (4)$$

Parameter T menyatakan jumlah daun pada pohon, w adalah bobot daun, sedangkan γ dan λ merupakan hiperparameter regularisasi untuk mengontrol kompleksitas model.

2.7 Evaluasi Model

Setiap model dievaluasi menggunakan empat metrik regresi pada data uji:

a. *Root Mean Squared Error (RMSE):*

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

b. *Mean Absolute Error (MAE):*

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (6)$$

c. *Mean Absolute Percentage Error (MAPE):*

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

d. *Koefisien Determinasi (R^2):*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

di mana n= jumlah sampel, p= jumlah fitur. Pendekatan evaluasi ini digunakan dalam penelitian regresi karena mampu merepresentasikan kinerja model secara komprehensif [29].

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Selain itu, model divalidasi menggunakan skema 10-fold cross-validation pada data latih dengan metrik RMSE untuk menilai stabilitas prediksi [22]. Nilai rata-rata RMSE merepresentasikan tingkat kesalahan umum model, sedangkan standar deviasi menunjukkan konsistensi error antar fold. Nilai MAPE perlu diinterpretasikan dengan mempertimbangkan skala target dan ukuran dataset. Tahap evaluasi menghasilkan metrik kinerja setiap konfigurasi model yang selanjutnya digunakan sebagai dasar perbandingan antar metode *feature selection*, sebelum dilakukan validasi akhir pada model terbaik [19].

2.8 *Comparison*

Tahap *comparison* dilakukan dengan membandingkan kinerja *XGBoost Regression* pada skenario *baseline*, *Forward Selection*, *Backward Elimination*, dan *Optimized Selection* menggunakan metrik RMSE, MAE, MAPE, dan R^2 pada data uji [18]. Perbandingan ini bertujuan menilai efektivitas setiap metode *feature selection* dalam meningkatkan kinerja prediksi dan stabilitas prediksi dibandingkan model tanpa seleksi fitur [30]. Hasil evaluasi komparatif tersebut digunakan sebagai dasar penentuan model terbaik [28].

2.9 *Final Model Validation*

Tahap final model validation diterapkan pada model terbaik hasil proses comparison untuk menguji kemampuan generalisasi model [17]. Validasi dilakukan menggunakan 10-Fold *Cross-Validation* guna memastikan konsistensi performa pada berbagai subset data [21]. Selain itu, visualisasi nilai aktual dan prediksi serta analisis *feature importance* digunakan untuk mendukung interpretabilitas model [29].

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Bagian ini menyajikan hasil eksperimen yang diperoleh dari penerapan algoritma *XGBoost Regression* dengan berbagai strategi *feature selection* dalam memprediksi panjang siklus menstruasi.

3.1.1 Deskripsi Dataset dan *Preprocessing*

Dataset terdiri dari 162 observasi dengan 13 variabel awal, meliputi 8 variabel numerik bertipe *integer* (number_of_peak, Age, Length_of_cycle, Estimated_day_of_ovulation, Length_of_Leutal_Phase, Length_of_menses, Income, Mean_of_length_of_cycle), 3 variabel numerik bertipe *float* (Weight, BMI, Menses_score), dan 2 variabel kategorikal bertipe *object* (Unusual_Bleeding, Height). Hasil pemeriksaan kualitas data menunjukkan bahwa tidak terdapat duplikasi observasi maupun missing value (null/NaN) dalam dataset, sehingga seluruh 162 observasi dapat dipertahankan tanpa memerlukan imputasi atau penghapusan baris.

Selanjutnya, kolom *Income* dihapus karena seluruh nilainya konstan (bernilai nol), sehingga tidak memberikan informasi yang relevan bagi model prediksi. Setelah penghapusan, dataset terdiri dari 12 variabel. Fitur kategorikal kemudian dikonversi ke bentuk numerik. Variabel *Unusual_Bleeding* yang memiliki dua kategori (‘yes’, ‘no’) dikodekan menggunakan *label encoding* biner (‘no’ = 0, ‘yes’ = 1). Variabel *Height* yang semula berbentuk teks (misalnya “5'3”, “5 6”) dikonversi ke dalam satuan sentimeter (cm) menggunakan aturan konversi standar (1 kaki = 30,48 cm; 1 inci = 2,54 cm), dan diberi nama Height_cm. Tidak terdapat nilai yang gagal dikonversi, sehingga tidak diperlukan imputasi median.

Identifikasi outlier menggunakan metode IQR menunjukkan terdapat beberapa nilai di luar rentang IQR pada variabel Weight (rentang: 40–85 kg) dan BMI (rentang: 16,2–32,9). Namun nilai-nilai tersebut masih berada dalam rentang fisiologis yang wajar untuk wanita usia 14–25 tahun, sehingga tidak dikategorikan sebagai outlier ekstrem dan tetap dipertahankan dalam analisis. Setelah preprocessing, dataset akhir terdiri dari 11 fitur numerik dan 1 variabel target, yaitu Length_of_cycle, yang merepresentasikan panjang siklus menstruasi dalam hari.

Dataset dibagi menjadi 129 sampel *training* (80%) dan 33 sampel *testing* (20%) secara acak dengan *random_state* tetap untuk memastikan reproduktibilitas. Seluruh fitur numerik dinormalisasi menggunakan *StandardScaler* dengan parameter fit pada data *training* untuk menghindari data *leakage*.

3.1.2 Hasil *Feature Selection*

Penelitian ini menerapkan tiga pendekatan *feature selection*: Forward Selection, Backward Elimination, dan Optimized Selection untuk mengidentifikasi subset fitur paling relevan terhadap prediksi panjang siklus menstruasi.

a. Hasil Seleksi Fitur dengan *Forward Selection*

Metode *Forward Selection* bekerja dengan menambahkan fitur secara bertahap ke dalam model berdasarkan kontribusinya terhadap peningkatan nilai R^2 pada data validasi. Proses ini dimulai dari subset fitur kosong dan dihentikan ketika penambahan fitur baru tidak lagi memberikan peningkatan performa yang substansial. Nilai yang ditampilkan pada kolom tabel berikut merepresentasikan skor kumulatif koefisien determinasi (R^2) yang diperoleh setelah setiap fitur ditambahkan secara bertahap ke dalam model.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Tabel 3. Hasil fitur Forward Selection

Feature	R^2 Cumulative Score
Mean_of_length_of_cycle	0,679079
Height_cm	0,768429
Length_of_menses	0,882572
Menses_score	0,899137
Age	0,900460

Hasil *Forward Selection* sebagaimana ditunjukkan pada Tabel 3 menghasilkan lima fitur terpilih, yaitu Mean_of_length_of_cycle, Height_cm, Length_of_menses, Menses_score, dan Age. Fitur *Mean_of_length_of_cycle* dipilih pertama karena memberikan kontribusi terbesar secara individual terhadap peningkatan performa model. Fitur-fitur berikutnya ditambahkan secara bertahap hingga model mencapai nilai R^2 sebesar 0,9005.

Hasil ini menunjukkan bahwa kombinasi lima fitur tersebut sudah mampu merepresentasikan variasi panjang siklus menstruasi secara optimal. Pengurangan jumlah fitur dari sebelas menjadi lima tidak hanya meningkatkan performa prediksi, tetapi juga menghasilkan model yang lebih ringkas dan mudah diinterpretasikan.

b. Hasil Seleksi Fitur dengan *Backward Elimination*

Berbeda dengan *Forward Selection*, metode *Backward Elimination* memulai proses seleksi dengan seluruh fitur yang tersedia, kemudian secara bertahap menghapus fitur dengan kontribusi terendah berdasarkan *feature importance*. Proses eliminasi dihentikan ketika penghapusan fitur menyebabkan penurunan nilai R^2 melebihi ambang batas toleransi yang telah ditetapkan. Nilai pada kolom tabel berikut merupakan XGBoost Feature Importance berbasis gain yang dihitung dari model final setelah proses eliminasi selesai. Gain mengukur peningkatan kualitas split yang dihasilkan suatu fitur dalam struktur pohon keputusan.

Tabel 4. Hasil fitur Backward Elimination

Feature	XGBoost Feature Importance (Gain)
BMI	0,092198
Menses_score	0,135520
Unusual_Bleeding	0,370372
Mean_of_length_of_cycle	0,401910

Hasil *Backward Elimination* pada Tabel 4 menunjukkan bahwa empat fitur yang dipertahankan adalah Mean_of_length_of_cycle, Unusual_Bleeding, Menses_score, dan BMI. Metode ini berhasil mengurangi jumlah fitur secara substansial, namun performa model yang dihasilkan tidak meningkat secara optimal dibandingkan *Forward Selection*. Hal ini mengindikasikan bahwa meskipun beberapa fitur memiliki kontribusi individual yang relatif kecil, interaksi antar fitur tertentu tetap berperan penting dalam meningkatkan performa prediksi.

c. Hasil Seleksi Fitur dengan *Optimized Selection*

Pendekatan *Optimized Selection* menggunakan strategi *ensemble feature selection* dengan mengombinasikan lima metode, yaitu korelasi Pearson, *mutual information*, F-test (ANOVA), *feature importance XGBoost*, dan *Recursive Feature Addition*. Setiap metode menghasilkan skor relevansi fitur yang kemudian dinormalisasi dan dirata-ratakan menjadi *ensemble score*. Nilai pada kolom tabel berikut merupakan Ensemble Feature Score, yaitu skor gabungan yang dihitung dari rata-rata terbobot lima metode seleksi fitur tersebut.

Tabel 5. Hasil Feature Selection Optimized Selection

Feature	Ensemble Feature Score
Mean_of_length_of_cycle	0,36801746
Unusual_Bleeding	0,26527977
Estimated_day_of_ovulation	0,11259647
BMI	0,10009686
Weight	0,09653846
Height_cm	0,057471033

Berdasarkan hasil pada Tabel 5 di atas, terdapat enam fitur terpilih yaitu Mean_of_length_of_cycle, Unusual_Bleeding, Estimated_day_of_ovulation, BMI, Weight, dan Height_cm. Keenam fitur tersebut kemudian digunakan sebagai input model untuk evaluasi kinerja. Pendekatan ini bersifat lebih komprehensif karena menilai relevansi fitur dari berbagai perspektif statistik dan berbasis model.

d. Hasil Ringkasan Fitur Terpilih

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

Tabel 6. Hasil Fitur Terpilih

Algoritma Pemodelan	Jumlah Fitur	Daftar Fitur Terpilih
Forward Selection	5	Mean_of_length_of_cycle, Height_cm, Length_of_menses, Menses_score, Age
Backward Elimination	4	Mean_of_length_of_cycle, Unusual_Bleeding, Menses_score, BMI
Optimized Selection	6	Mean_of_length_of_cycle, Unusual_Bleeding, Estimated_day_of_ovulation, BMI, Weight, Height_cm

Tabel 6 menyajikan ringkasan fitur yang terpilih dari masing-masing metode *feature selection*. Terlihat bahwa fitur Mean_of_length_of_cycle muncul secara konsisten di ketiga metode, menunjukkan relevansi dominannya dalam memprediksi panjang siklus menstruasi.

3.1.3 Hasil Evaluasi *XGBoost Regression* per Skenario

Setiap *subset* fitur hasil seleksi selanjutnya digunakan untuk melatih model *XGBoost Regression*. Perbandingan kinerja seluruh model, termasuk baseline, dilakukan menggunakan metrik RMSE, MAE, MAPE, dan R^2 pada data *testing*, serta evaluasi 10-fold *cross-validation* dengan metrik RMSE pada data *training* untuk menilai stabilitas model.

Tabel 7. Hasil Evaluasi Performa Model

Algoritma Pemodelan	RMSE	MAE	MAPE (%)	R^2	CV RMSE (Mean ± Std)
XGBoost Baseline	2,3892	1,0857	3,58	0,7309	1,9763 ± 0,9657
Forward Selection + XGBoost	1,4531	0,5707	1,73	0,9005	2,1276 ± 0,9255
Backward Elimination + XGBoost	2,1902	1,3592	4,47	0,7739	2,1774 ± 0,7897
Optimized Selection + XGBoost	1,9676	1,1962	3,95	0,8175	2,1546 ± 0,7528

Hasil evaluasi menunjukkan Forward Selection + XGBoost menghasilkan performa terbaik dengan R^2 = 0,9005 dan RMSE = 1,4531 hari, diikuti Optimized Selection (R^2 = 0,8175, RMSE = 1,9676 hari), Backward Elimination (R^2 = 0,7739, RMSE = 2,1902 hari), dan model baseline (R^2 = 0,7309, RMSE = 2,3892 hari). Evaluasi stabilitas melalui CV RMSE perlu dibandingkan dengan RMSE testing guna memastikan kemampuan generalisasi model. Baseline memiliki CV RMSE terendah (1,9763 ± 0,9657), tetapi RMSE testing tertinggi (2,3892), menunjukkan penurunan performa pada data uji. Forward Selection memiliki CV RMSE 2,1276 ± 0,9255 dengan RMSE testing lebih rendah (1,4531), mengindikasikan generalisasi yang baik tanpa *overfitting*. Backward Elimination (2,1774 ± 0,7897) dan Optimized Selection (2,1546 ± 0,7528) menunjukkan konsistensi antar-fold (std < 0,80), namun akurasi testing tetap di bawah Forward Selection. Secara keseluruhan, meskipun baseline memiliki CV RMSE lebih kecil, Forward Selection memberikan keseimbangan terbaik antara akurasi dan stabilitas.

3.1.4 Visualisasi Prediksi dan Residual

Berdasarkan hasil evaluasi, model Forward Selection + XGBoost ditetapkan sebagai model terbaik dan selanjutnya divalidasi pada data uji.

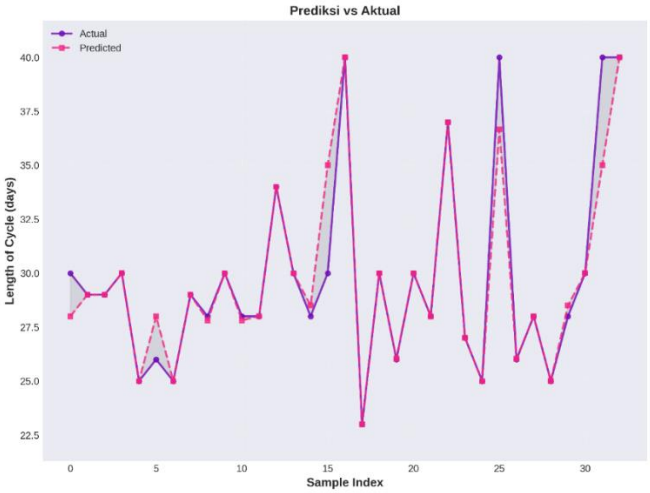
Tabel 8. Hasil Prediksi dari Fitur Terbaik

Actual	Predicted
30	27,999838
29	29,000227
29	29,000470
30	30,003000
25	25,000120
26	27,999838
25	25,000708
29	28,999653
28	27,813007
30	30,000916

Tabel 8 menyajikan perbandingan antara nilai aktual dan nilai prediksi. Secara umum, nilai prediksi sangat mendekati nilai aktual, dengan selisih yang relatif kecil. Sebagai contoh, nilai aktual 30 hari diprediksi sebesar 27,99–30,00 hari, sedangkan nilai aktual 29 hari diprediksi sekitar 29,00 hari. Deviasi maksimum yang

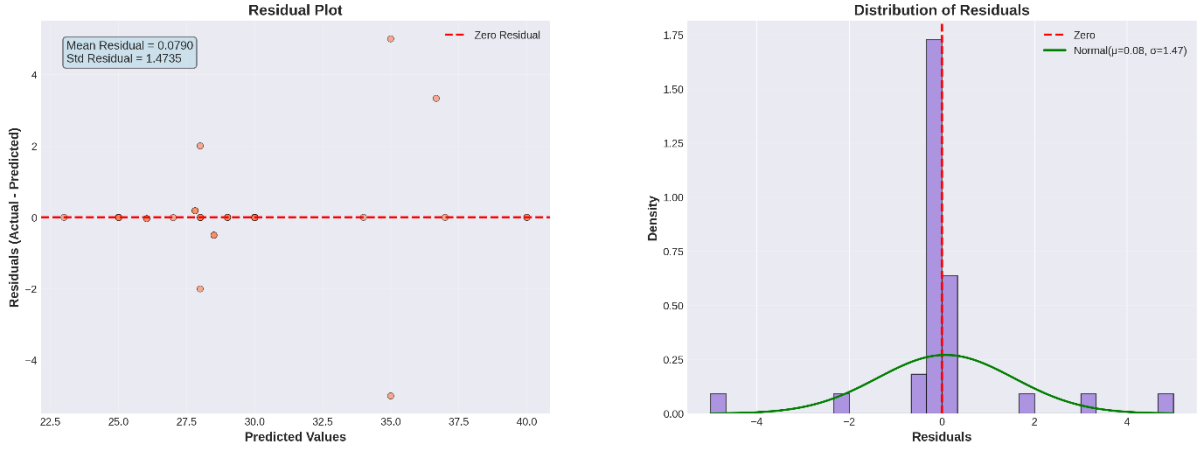
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

terlihat masih berada dalam rentang ±2 hari, yang secara klinis tergolong kecil dan dapat diterima untuk konteks pemantauan siklus menstruasi.



Gambar 2. Perbandingan Nilai Aktual dan Prediksi

Pada gambar 2 memperlihatkan perbandingan nilai aktual dan prediksi secara visual. Kurva prediksi mengikuti pola kurva aktual secara konsisten di seluruh 33 data uji, dengan jarak antar kurva yang sangat kecil, mengonfirmasi rendahnya tingkat kesalahan model.



Gambar 3. Visualisasi Prediksi dan Residual

Gambar 3 menyajikan analisis residual dalam dua panel yang saling melengkapi. Panel kiri (*Residual Plot*) memperlihatkan residual tersebar acak di sekitar garis nol (*mean* = 0,079) tanpa membentuk pola corong maupun kurva sistematis, mengindikasikan tidak adanya heteroskedastisitas. Panel kanan (*Distribution of Residuals*) menunjukkan histogram residual yang mendekati distribusi normal dengan kurva *fitting* N(0,08; 1,47), menunjukkan kecenderungan mendekati distribusi normal. Secara bersama, kedua panel menunjukkan bahwa *error* prediksi terpusat di sekitar nol, tersebar merata, dan tidak mengandung bias sistematis.

3.2 Pembahasan

3.2.1 Perbandingan *Baseline* vs *Feature Selection*

Berdasarkan hasil evaluasi pada Tabel 7, model baseline menghasilkan R^2 = 0,7309 dengan RMSE = 2,3892 hari, menunjukkan kemampuan cukup baik dalam menangkap pola hubungan antara variabel input dan panjang siklus menstruasi, dengan kemampuan menjelaskan sekitar 73% variasi. Namun, penggunaan seluruh fitur tanpa seleksi berpotensi memasukkan fitur tidak relevan atau redundan, sehingga menimbulkan noise dan menurunkan ketepatan prediksi.

Forward Selection meningkatkan R^2 dari 0,7309 menjadi 0,9005 (peningkatan 23,2%) dengan reduksi fitur dari 11 menjadi 5 (pengurangan 54,5%). Pengurangan fitur secara selektif mengurangi noise dan memperkuat hubungan antara fitur dan variabel target. Backward Elimination menghasilkan R^2 = 0,7739 (+5,9% vs baseline) dengan reduksi fitur hingga 63,6% (4 fitur), namun performa tidak optimal karena error relatif lebih besar pada nilai ekstrem. Optimized Selection mencapai R^2 = 0,8175 (+11,9%) dengan 6 fitur dan RMSE 1,9676. Pendekatan ensemble yang menggabungkan lima metode seleksi memang meningkatkan stabilitas pemilihan variabel, namun

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

agregasi skor berbasis rata-rata cenderung mempertahankan fitur yang tidak memberikan kontribusi prediktif tambahan. Akibatnya, meskipun jumlah fitur lebih banyak, informasi yang diperoleh tidak meningkat secara proporsional, sehingga penurunan RMSE terbatas dan tetap lebih tinggi dibanding Forward Selection (1,4531). Hal ini menunjukkan bahwa stabilitas seleksi tidak selalu berbanding lurus dengan peningkatan akurasi prediktif.

Analisis gap CV-Testing pada Tabel 7 menunjukkan Baseline memiliki CV RMSE terendah (1,9763) namun gap positif tertinggi (+0,41), mengindikasikan penurunan kemampuan generalisasi pada data uji. Forward Selection memiliki CV RMSE 2,1276 dengan gap negatif (-0,67), menunjukkan generalisasi yang baik tanpa indikasi overfitting. Backward Elimination memiliki konsistensi tertinggi (std 0,7897) dengan gap terkecil (+0,01). Optimized Selection memiliki gap -0,19, menunjukkan generalisasi baik namun tidak seoptimal Forward Selection.

3.2.2 Analisis Penyebab dan Perbedaan Performa

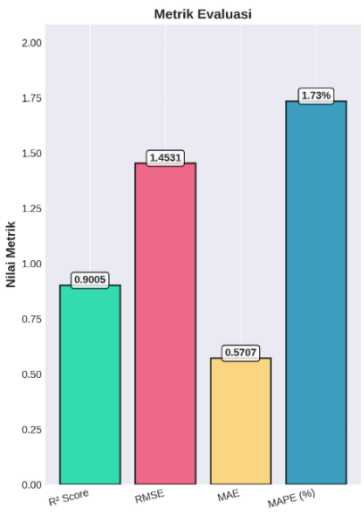
Keunggulan Forward Selection terletak pada strategi incremental yang memastikan setiap fitur memberikan kontribusi nyata. Berdasarkan Tabel 3, Mean_of_length_of_cycle yang dipilih pertama berkontribusi R^2 0,6791 (67,9% variasi), kemudian kombinasi dengan Height_cm meningkatkan R^2 menjadi 0,7684 (+14,7%), menunjukkan interaksi sinergis antara pola historis siklus dengan faktor antropometri yang mempengaruhi metabolisme hormonal. Penambahan Length_of_menses meningkatkan R^2 menjadi 0,8826 (+11,4%), mengindikasikan durasi menstruasi berperan penting secara fisiologis.

Keterbatasan Backward Elimination terlihat dari kehilangan informasi interaktif akibat eliminasi fitur. Eliminasi Height_cm, Length_of_menses, dan Age menghilangkan informasi interaksi fitur yang kompleks. Proses eliminasi dihentikan prematur ketika penghapusan BMI (importance terendah 0,0922) menyebabkan penurunan CV R^2 sebesar 11,17%, jauh melebihi threshold 5%.

Trade-off metode ini adalah efisiensi reduksi fitur tinggi namun risiko kehilangan informasi interaktif. Trade-off Optimized Selection terlihat dari ensemble scoring yang tidak selalu mencerminkan kontribusi optimal. Meskipun menggunakan 6 fitur, penambahan Weight dan Estimated_day_of_ovulation tidak memberikan kontribusi substansial terhadap peningkatan R^2. Weight dan BMI memiliki korelasi tinggi ($r > 0,8$), menciptakan redundansi tanpa peningkatan performa. Standar deviasi CV (0,7528) mengindikasikan sensitivitas terhadap subset data training berbeda.

3.2.3 Analisis *Error* dan Interpretasi Fitur

Berdasarkan hasil evaluasi, model Forward Selection + XGBoost menunjukkan performa prediksi yang sangat baik. Gambar 4 menyajikan visualisasi metrik kinerja model terbaik.



Gambar 4. Metrik Kinerja Model *Forward Selection*

Nilai MAE sebesar 0,5707 hari menunjukkan bahwa rata-rata kesalahan absolut prediksi hanya sekitar setengah hari. Artinya, secara rata-rata model meleset kurang dari 14 jam dari nilai aktual, yang mencerminkan ketepatan prediksi yang sangat tinggi. RMSE sebesar 1,4531 hari sedikit lebih besar dibandingkan MAE karena RMSE memberikan penalti lebih besar pada kesalahan yang relatif lebih tinggi. Pada tabel, terdapat beberapa prediksi dengan selisih mendekati 2 hari, seperti aktual 30 hari yang diprediksi sekitar 27,99 hari atau aktual 26 hari yang diprediksi sekitar 27,99 hari. Meskipun jumlah kasus ini terbatas, keberadaan error yang lebih besar tersebut meningkatkan nilai RMSE. Namun, karena error ekstrem jarang terjadi, nilai RMSE tetap berada pada kisaran rendah.

Secara keseluruhan, distribusi error model menunjukkan tiga pola utama: (1) mayoritas prediksi memiliki deviasi di bawah 1 hari dan hanya sebagian kecil mendekati ±2 hari; (2) error yang lebih besar sesekali muncul

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

pada siklus >35 hari namun tidak membentuk pola sistematis; dan (3) residual terdistribusi simetris di sekitar nol tanpa kecenderungan overestimation maupun underestimation, sebagaimana dikonfirmasi oleh mean residual 0,079 dan std 1,4735 pada Gambar 3.

MAPE sebesar 1,73% menunjukkan bahwa kesalahan prediksi relatif terhadap nilai aktual sangat kecil secara persentase. Dengan rata-rata panjang siklus sekitar 25–30 hari, selisih ±1 hari hanya merepresentasikan error sekitar 3–4%, dan karena mayoritas error berada jauh di bawah 1 hari, MAPE turun hingga di bawah 2%. Hal ini menegaskan bahwa model memiliki tingkat kesalahan relatif yang sangat rendah dan konsisten di berbagai nilai siklus. Sementara itu, R^2 sebesar 0,9005 menunjukkan bahwa model mampu menjelaskan sekitar 90% variasi panjang siklus menstruasi berdasarkan fitur-fitur terpilih.

Fitur Mean_of_length_of_cycle muncul secara konsisten pada seluruh metode seleksi, mengindikasikan peran dominannya dalam pemodelan. Sebagai representasi pola historis individu, fitur ini secara statistik dan fisiologis memiliki hubungan yang kuat dengan panjang siklus berikutnya.

3.2.4 Perbandingan dengan Penelitian Terdahulu

Temuan ini sejalan dengan penelitian sebelumnya yang menekankan pentingnya data historis siklus menstruasi dalam pemodelan prediktif [6], [8], namun penelitian ini melangkah lebih jauh dengan membandingkan secara sistematis berbagai strategi *feature selection*. Meskipun Rego [6] dan Khairunisa dkk. [8] menunjukkan keunggulan model berbasis *ensemble* seperti *XGBoost Regression*, hasil penelitian ini memberikan bukti empiris bahwa integrasi *XGBoost Regression* dengan *forward selection* meningkatkan performa secara substansial dibanding model *baseline* (R^2 dari 0,7309 menjadi 0,9005) maupun metode seleksi fitur lain. Dengan $R^2 > 0,90$, performa model ini juga lebih tinggi dibandingkan laporan sebelumnya yang umumnya berada pada kisaran 0,70–0,85 [8], [9], [10], memperkuat potensi penerapannya dalam sistem digital untuk pemantauan kesehatan reproduksi perempuan.

3.2.5 Keterbatasan dan Implikasi

Penelitian ini memiliki beberapa keterbatasan. Ukuran dataset relatif kecil (n = 162; 129 *training*, 33 *testing*) sehingga berpotensi membatasi generalisasi pada populasi yang lebih luas. Representasi sampel yang homogen dengan rentang usia 14–25 tahun juga tidak mencakup kelompok usia lain maupun kondisi medis seperti PCOS atau endometriosis. Selain itu, fitur yang digunakan masih terbatas pada data antropometri dan riwayat siklus, tanpa memasukkan faktor *lifestyle* (stres, aktivitas fisik).

Meskipun demikian, model Forward Selection + XGBoost dengan R^2 = 0,9005 dan MAE = 0,57 hari menunjukkan potensi implementasi dalam aplikasi *mobile health* untuk pemantauan siklus menstruasi. Deviasi prediksi > 2 hari (6,1% kasus) dapat dimanfaatkan sebagai indikator konsultasi medis dan sistem *early warning* terhadap irregularitas. Dominasi fitur Mean_of_length_of_cycle (kontribusi 67,9%) menegaskan pentingnya pola historis individu dalam prediksi personal. Penelitian selanjutnya disarankan menggunakan dataset lebih besar (n > 500) dengan variasi demografis lebih luas, integrasi fitur hormonal dan *lifestyle*, perbandingan dengan metode *deep learning* (LSTM, GRU) untuk menangkap *temporal dependencies*, serta validasi eksternal pada populasi independen guna menguji generalisasi model.

4. KESIMPULAN

Penelitian ini mengevaluasi secara komparatif penerapan berbagai metode *feature selection* pada model *XGBoost Regression* untuk memprediksi panjang siklus menstruasi. Hasil penelitian menunjukkan bahwa penerapan seleksi fitur mampu meningkatkan performa prediksi dan stabilitas model regresi dibandingkan pendekatan tanpa seleksi fitur. Di antara metode yang diuji, *Forward Selection* memberikan performa terbaik dengan nilai R^2 sebesar 0,9005, serta kesalahan prediksi terendah dengan RMSE 1,45 hari, MAE 0,57 hari, dan MAPE 1,73%, yang mencerminkan ketepatan prediksi yang tinggi. Keunggulan ini diperoleh melalui pemilihan fitur paling relevan secara bertahap sehingga menghasilkan model yang lebih sederhana dan efisien. Temuan ini menegaskan pentingnya strategi seleksi fitur dalam pengembangan model regresi prediktif di bidang kesehatan reproduksi perempuan. Meskipun demikian, mengingat ukuran dataset yang terbatas dan karakteristik sampel yang relatif homogen, hasil ini masih memerlukan validasi pada populasi yang lebih luas sebelum diimplementasikan secara klinis. Oleh karena itu, untuk penelitian selanjutnya disarankan menggunakan dataset lebih besar (n > 500) dengan variasi demografis lebih luas (usia 18-45 tahun, kondisi PCOS/endometriosis), integrasi fitur hormonal dan *lifestyle* (stres, pola tidur, aktivitas fisik), serta eksplorasi metode *deep learning* (LSTM, GRU) untuk menangkap *temporal dependencies* guna meningkatkan prediksi pada siklus menstruasi.

REFERENCES

[1] J. Le Yu *et al.*, "Tracking of menstrual cycles and prediction of the fertile window via measurements of basal body temperature and heart rate as well as machine-learning algorithms," *Reprod. Biol. Endocrinol.*, vol. 20, no. 1, pp. 1–

12, 2022, doi: 10.1186/s12958-022-00993-4.

[2] W. H. Organization, *Sexual and Reproductive Health and Rights*. WHO Press, 2021.

[3] F. Ismawati and dr Adhy Purnawan MKes, "Hubungan Aktivitas Fisik Dan Tingkat Stres Dengan Siklus Menstruasi Pada Siswi Smk Informatika Ciputat Tahun 2022," *Fram. Heal. J.*, vol. 1, no. 2, pp. 173–180, 2022.

[4] G. Kilungeja, K. Graham, X. Liu, and M. Nasseri, "Machine learning-based menstrual phase identi fi cation using wearable device data Check for updates," *npj Women's Heal.*, pp. 1–10, 2025, doi: 10.1038/s44294-025-00078-8.

[5] S. Aggarwal, *Machine Learning for Healthcare Applications*. Springer, 2022.

[6] R. C. B. Rego, "Predictive Modeling of Menstrual Cycle Length: A Time Series Forecasting Approach," pp. 1–12, 2023, [Online]. Available: http://arxiv.org/abs/2308.07927

[7] D. Reskianto and M. A. Barata, "Forecasting Metode Single Exponential Smoothing Dalam Meramalkan Penjualan Barang," pp. 435–444, 2023.

[8] M. Khairunisa, D. Made, S. Amanda, I. G. Ngurah, and L. Wijayakusuma, "Perbandingan Metode Machine Learning untuk Analisis dan Prediksi Siklus Menstruasi," vol. 08, pp. 111–115, 2024.

[9] Tanmay Thakur, Saurabh Kadam, Nikita Patil, and Chinmayee Achrekar, "Machine Learning in Period, Fertility and Ovulation Tracking Application," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 4, pp. 200–203, 2023, doi: 10.48175/ijarsct-9286.

[10] O. J. C. *et al.*, "Improving Menstrual Cycle Prediction Accuracy using Advanced Machine Learning Model Methods," *J. IoT Mach. Learn.*, vol. 1, no. 2, pp. 1–7, 2023, doi: 10.48001/joitml.2023.121-7.

[11] M. B. Prayogi, F. Apriani, U. N. Huda, and I. Pendidikan, "PREDIKSI ANGKA HARAPAN HIDUP MENGGUNAKAN RANDOM," vol. 2, no. 1, pp. 112–121, 2025.

[12] M. Khairunisa, D. Made, S. Amanda, I. G. Ngurah, and L. Wijayakusuma, "Comparison of Machine Learning Methods for Menstrual Cycle Analysis and Prediction," vol. 9, no. 2, pp. 348–353, 2025.

[13] M. F. Asnawi, H. H. Bisono, and M. A. Megantara, "Aplikasi Prediksi Banjir Menggunakan Algoritma XGBoost Berbasis Website," vol. 7, no. 2, pp. 379–389, 2024.

[14] M. R. Andryan, M. Fajri, T. Informatika, U. S. Karawang, T. Timur, and J. Barat, "Komparasi kinerja algoritma xgboost dan algoritma support vector machine (svm) untuk diagnosa penyakit kanker payudara," vol. 6, no. 1, pp. 1–5, 2022.

[15] A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, "PREDIKSI KUALITAS UDARA MENGGUNAKAN XGBOOST DENGAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) BERDASARKAN INDEKS STANDAR PENCEMARAN UDARA (ISPU)," vol. 7, no. 1, pp. 214–219, 2023.

[16] A. Rahmat, D. W. Utomo, U. D. Nuswantoro, and U. D. Nuswantoro, "DIABETES DETECTION USING STACKING TECHNIQUE : A COMBINATION OF XGBOOST , GRADIENT BOOSTING , AND META MODEL," vol. 10, no. 2, pp. 912–921, 2025.

[17] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," vol. 7, no. 3, pp. 193–198, 2022.

[18] F. Kamalov, S. Elnaffar, A. Cherukuri, and A. Jonnalagadda, "Forward Feature Selection: Empirical Analysis," *J. Intell. Syst. Internet Things*, vol. 11, no. 01, pp. 44–54, 2024.

[19] F. Dardiri, "Comparison Of Naïve Bayes and Decision Trees in Determining the Best Manager of Nurul Jadid Islamic Boarding School Based on Forward Selection," vol. 8, no. 2, pp. 689–698, 2024.

[20] J. Melvin and A. Soraya, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," vol. 2, no. 2, 2023.

[21] K. Chotchantarakun, "Optimizing Sequential Forward Selection on Classification Using Genetic Algorithm Related work Feature selection process," vol. 47, pp. 81–90, 2023.

[22] D. Datasets, F. I. Silfana, and M. A. Barata, "Using K-NN Algorithm for Evaluating Feature Selection on High," vol. 17, no. 2, 2024.

[23] S. D. Amalia, M. A. Barata, and P. E. Yuwita, "Optimization of Random Forest Algorithm with Backward Elimination Method in Classification of Academic Stress Levels," vol. 9, no. 3, 2025.

[24] U. D. Nuswntoro, "C4 . 5 Forward Selection Based Algorithm for Class Level Classification of Nurul Jadid Islamic Boarding School Students," vol. 8, no. 2, pp. 699–712, 2024.

[25] S. Santhi and S. S. Nidhyananthan, "OPTIMIZED FEATURE SELECTION FOR PCOS DISEASE," no. 9, pp. 1430–1439, 2023, doi: 10.7546/CRABS.2023.09.14.

[26] R. A. Saputra *et al.*, "Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 61–67, 2020, doi: 10.1088/1742-6596/1641/1/012025.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9526

[27] M. Rai, “Menstrual Health and PCOD Risk Detection Dataset,” 2025.

[28] S. G. Fashoto, E. Mbunge, and G. Ogunleye, “IMPLEMENTATION OF MACHINE LEARNING FOR PREDICTING MAIZE CROP YIELDS USING MULTIPLE,” vol. 6, no. 1, pp. 679–697, 2021.

[29] I. R. Dina, M. A. Barata, and P. E. Yuwita, “Penerapan Data Mining pada Algoritma Multiple Linear Regression dalam Peramalan Harga Emas,” vol. 11, no. 1, pp. 1–7, 2025.

[30] C. Series, “Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease”, doi: 10.1088/1742-6596/1511/1/012001.