

Analisis Komparasi Random Forest dan XGBoost untuk Prediksi Produksi Kelapa Sawit Berbasis SHAP

Random Forest and XGBoost Comparison for SHAP Based Palm Oil Production Prediction

Risfi Ayu Sandika*, Chandra Kusuma Dewa

Fakultas Teknologi Industri, Program Studi Informatika, Universitas Islam Indonesia, Yogyakarta, Indonesia

Email: ¹*24917035@students.uui.ac.id, ²135231101@uui.ac.id

Email Penulis Korespondensi: 24917035@students.uui.ac.id

Received: 10/08/2026, Revised: 24/08/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstrak—Produksi kelapa sawit dipengaruhi oleh berbagai faktor, seperti umur tanaman, luas lahan, jumlah pokok produktif, dan penggunaan pupuk. Fluktuasi produksi yang terjadi menyebabkan perusahaan mengalami kesulitan dalam menyusun perencanaan produksi, distribusi, dan strategi pemupukan secara optimal. Berbeda dari penelitian sebelumnya, penelitian ini menggunakan dataset yang diperoleh secara langsung dari perusahaan untuk membandingkan kinerja algoritma Random Forest dan XGBoost dalam memprediksi produksi kelapa sawit, sekaligus menganalisis variabel mana yang paling memengaruhi hasil prediksi menggunakan SHAP (SHapley Additive exPlanations). Data yang digunakan merupakan data historis produksi kelapa sawit PT XYZ selama periode Januari 2020 hingga Desember 2025 sebanyak 588 data dengan pendekatan multivariate time series. Tahapan penelitian meliputi preprocessing data, pembentukan sliding window dengan sequence length 3 dan 6 bulan, pelatihan model, hyperparameter tuning, evaluasi menggunakan Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan Coefficient of Determination (R^2), serta interpretasi model menggunakan SHAP. Hasil penelitian menunjukkan bahwa kedua model memperoleh performa terbaik pada sequence length 3 bulan. Model XGBoost memberikan hasil yang lebih baik dibandingkan Random Forest dengan nilai MAE sebesar 553.293, RMSE sebesar 1.496.990, dan R^2 sebesar 0,877, sedangkan Random Forest memperoleh MAE sebesar 1.283.547, RMSE sebesar 1.792.539, dan R^2 sebesar 0,824. Analisis SHAP menunjukkan bahwa bulan merupakan faktor yang paling berpengaruh terhadap prediksi produksi kelapa sawit, diikuti oleh lag_1, lag_3 dan luas. Penelitian ini memberikan kontribusi yaitu perbandingan random forest dan XGBoost pada data produksi kelapa sawit primer yang diperoleh langsung dari perusahaan, serta interpretasi SHAP yang dapat dijadikan acuan praktis bagi PT XYZ dalam perencanaan produksi dan strategi perusahaan. Hasil penelitian ini menunjukkan bahwa kombinasi XGBoost dan SHAP mampu menghasilkan prediksi yang akurat sekaligus memberikan interpretasi terhadap faktor-faktor yang memengaruhi produksi sehingga dapat mendukung pengambilan keputusan dalam pengelolaan produksi kelapa sawit.

Kata Kunci: Kelapa Sawit; Prediksi Produksi; Random Forest; XGBoost; SHAP; Multivariate Time Series

Abstract—Palm oil production is influenced by various factors, such as plant age, land area, the number of productive trees, and fertilizer usage. Production fluctuations make it difficult for companies to formulate optimal plans for production, distribution, and fertilization strategies. Unlike previous studies, this research utilizes a dataset obtained directly from the company to compare the performance of Random Forest and XGBoost algorithms in predicting palm oil production, while also analyzing which variables most significantly influence prediction results using SHAP (SHapley Additive exPlanations). The study employs historical palm oil production data from PT XYZ covering the period from January 2020 to December 2025 comprising 588 data points using a multivariate time series approach. The research stages include data preprocessing, the creation of sliding windows with sequence lengths of 3 and 6 months, model training, hyperparameter tuning, evaluation using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2), as well as model interpretation using SHAP. The results indicate that both models achieved their best performance with a sequence length of 3 months. The XGBoost model outperformed Random Forest, achieving an MAE of 553,293, an RMSE of 1,496,990, and an R^2 of 0.877, whereas Random Forest recorded an MAE of 1,283,547, an RMSE of 1,792,539, and an R^2 of 0.824. SHAP analysis revealed that month was the most influential factor in predicting palm oil production, followed by lag_1, lag_3, and land area. This research contributes a comparison of Random Forest and XGBoost on primary palm oil production data obtained directly from the company, as well as a SHAP-based interpretation that can serve as a practical reference for PT XYZ in production planning and corporate strategy. These findings demonstrate that the combination of XGBoost and SHAP can generate accurate predictions while providing insights into the factors influencing production, thereby supporting decision-making in palm oil production management.

Keywords: Palm Oil; Production Prediction; Random Forest; XGBoost; SHAP; Multivariate Time Series

1. PENDAHULUAN

Salah satu sektor utama dalam perekonomian Indonesia adalah industri kelapa sawit [1]. Kelapa sawit adalah tanaman yang menghasilkan minyak nabati tinggi dibanding tanaman penghasil minyak nabati lainnya [2], menjadikannya sebagai sumber devisa terbesar sekaligus penyerap tenaga kerja dalam jumlah besar dan berbagai produk menjadikan bahan baku seperti bahan kimia, bahan pangan hingga sumber energi terbarukan [3] [4]. Permintaan pasar global yang terus meningkat memerlukan pemantauan yang ketat terhadap efisiensi produksi agar target produksi tercapai [3]. Salah satu perusahaan yang bergerak di bidang produksi kelapa sawit dengan skala besar yaitu PT XYZ yang berada di provinsi Riau, yang menjadi objek pada penelitian ini. Namun fakta dilapangan pengelolaan produksi kelapa sawit menghadapi berbagai tantangan yaitu serangan hama, penggunaan pupuk, cuaca hingga faktor usia tanaman [5]. Keempat faktor yang menjadi fokus penelitian ini yaitu umur tanaman, luas lahan, jumlah pokok produktif dan

penggunaan pupuk, saling berinteraksi secara kompleks dan tidak selalu bersifat linear terhadap hasil produksi yang diukur dalam kilogram atau ton per hektar [6], sehingga pendekatan analisis konvensional sulit menangkap pola hubungan secara akurat. Ketidakseimbangan hasil panen dengan permintaan pasar mempengaruhi dari distribusi dan bisnis. Kemudian dapat juga menyebabkan kelebihan kapasitas atau penurunan kapasitas, meningkatkan biaya operasional, hingga terjadi penurunan produktivitas tanaman. Kompleksitas hubungan antar variabel inilah yang mendorong kebutuhan terhadap pendekatan machine learning, yang mampu memodelkan pola non-linear antar faktor produksi secara lebih baik dibandingkan metode statistik konvensional. Salah satu faktor yang mempengaruhi pertumbuhan tanaman yaitu ketersediaan unsur hara [7]. Jadi ketersediaan unsur hara mempengaruhi produksi yang dihasilkan kelapa sawit. Kecukupan unsur hara yaitu keadaan dimana konsentrasi unsur hara esensial tersedia dalam jumlah yang cukup dalam mendukung pertumbuhan tanaman [8]. Unsur hara yang dibutuhkan oleh tanaman yaitu proses pemupukan. Pemupukan menjadi komponen penting dalam pertumbuhan kelapa sawit. Jadi untuk menjaga potensi hasil produksi kelapa sawit yang optimal, pemupukan adalah hal mutlak yang dilakukan [7]. Oleh karena itu, diperlukannya analisis kuantitatif yang mampu mengukur kontribusi variabel pupuk maupun faktor-faktor lainnya seperti umur tanaman, luas lahan, dan jumlah pokok produktif secara bersamaan terhadap hasil produksi, sehingga perusahaan dapat menyusun strategi pemupukan dan perencanaan produksi lebih efektif.

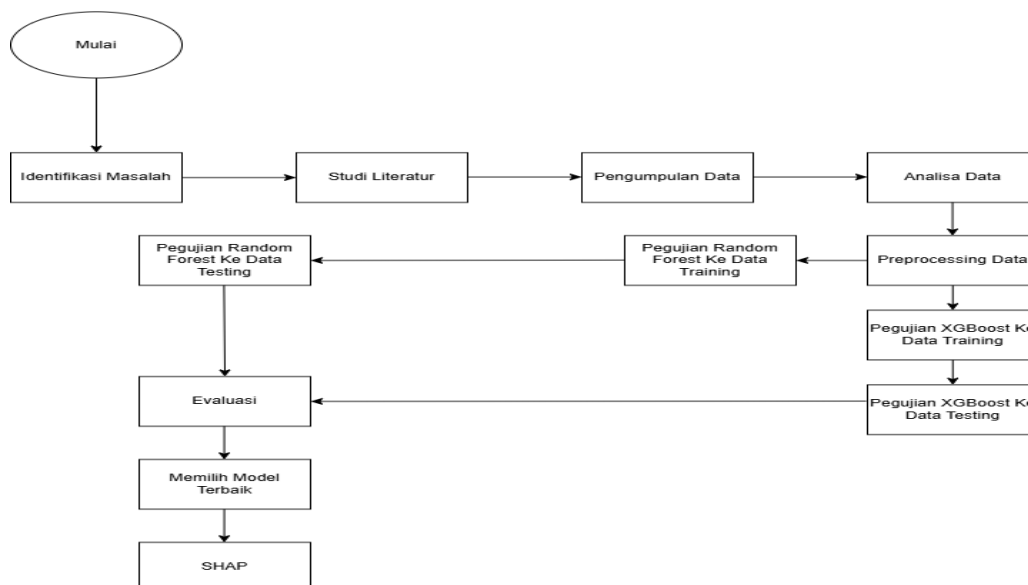
Berdasarkan penelitian sebelumnya mengenai prediksi penyakit gagal jantung dengan menggunakan random forest. Hasil penelitian metode random forest sukses meraih tingkat akurasi sebesar 82,608% dan dioptimalisasi dengan K-Fold serta GridSearch CV menjadi 85,058% [9]. Penelitian selanjutnya yaitu performance analysis of LSTM and XGBoost models optimization in forecasting crude palm oil (CPO) production at palm oil mill (POM). Penelitian tersebut membandingkan LSTM dan XGBoost untuk memprediksi produksi minyak CPO, sebelum dilakukannya hyperparameter tuning model XGBoost menghasilkan prediksi yang lebih unggul dengan data aktual dibanding LSTM. Setelah dilakukan hyperparameter tuning penelitian menunjukan LSTM memperoleh performa baik dengan nilai RMSE sebesar 21,04, MAPE 0,063, dan tingkat akurasi 93,7%, dibandingkan XGBoost yang memperoleh RMSE 22,17, MAPE 0,072, dan akurasi 92,8% [10]. Temuan ini menunjukan bahwa keunggulan XGBoost tidak bersifat mutlak dan sangat bergantung dengan karakteristik dataset serta proses tuning yang dilakukan, sehingga perbandingan langsung dengan algoritma lain seperti random forest tetap diperlukan untuk memastikan pemilihan model yang tepat pada kasus spesifik PT XYZ. Penelitian lainnya yaitu model prediksi produksi pertanian berbasis *machine learning* dan data lapangan menggunakan empat metode yaitu random forest, support vector regression, XGBoost, dan artificial neural network. Hasil penelitian menunjukan bahwa XGBoost memberikan performa terbaik dengan nilai R^2 sebesar 0,89 dan RMSE 0,52 ton/ha pada dataset pengujian [11]. Selanjutnya penelitian tentang practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Penelitian tersebut menyimpulkan bahwa SHAP meningkatkan transparansi dan kepercayaan terhadap model prediksi dengan memberikan informasi pengaruh setiap fitur terhadap hasil prediksi [12]. Meskipun demikian, keempat penelitian tersebut memiliki keterbatasan yang relevan bagi penelitian ini. Pertama, penelitian ini membandingkan random forest dan XGBoost dengan interpretasi SHAP secara bersamaan pada kasus produksi kelapa sawit dengan sumber data primer dari perusahaan. Kedua, tidak ada dari penelitian tersebut melakukan perbandingan performa random forest dan XGBoost dengan interpretasi SHAP secara bersamaan, sehingga mengetahui faktor yang mempengaruhi hasil prediksi pada konteks ini belum teridentifikasi. Penelitian ini mengisi kekosongan tersebut dengan membandingkan random forest dan XGBoost menggunakan data historis produksi kelapa sawit yang diperoleh dari PT XYZ, sekaligus menerapkan SHAP untuk identifikasi variabel yang berpengaruh terhadap hasil prediksi sangat efektif sebagai pertimbangan dalam pengambilan keputusan dan strategi produksi kelapa sawit lebih optimal [11].

Penelitian ini membandingkan model random forest dan XGBoost karena model tersebut merupakan algoritma machine learning berbasis decision tree yang banyak digunakan untuk permasalahan prediksi dan regresi. Meskipun keduanya menggunakan pendekatan yang sama, tetapi metode keduanya memiliki mekanisme teknik yang berbeda yaitu bagging dan boosting. Penelitian ini menggunakan dataset yang diperoleh secara langsung dari perusahaan dan tidak hanya membandingkan kinerja metode prediksi, tetapi juga menganalisis variabel yang paling berpengaruh terhadap hasil produksi menggunakan pendekatan SHAP. Data yang akan digunakan pada penelitian ini adalah data historis 6 tahun dengan multivariate time series. Penelitian ini menggunakan MAPE (*Mean Absolute Percentage Error*), MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*) dan R^2 (*Coefficient of Determination*). Dengan demikian penelitian ini diharapkan dapat menghasilkan prediksi yang akurat sekaligus interpretatif, sehingga memudahkan PT XYZ dalam menentukan perencanaan, pengawasan dan kebijakan secara tepat, serta meminimalkan resiko kerugian akibat ketidakpastian produksi [13]. Kontribusi utama penelitian ini adalah membandingkan performa random forest dan XGBoost berbasis data primer produksi kelapa sawit yang belum dilakukan pada penelitian sebelumnya, menerapkan interpretasi SHAP untuk mengidentifikasi variabel yang paling berpengaruh dan menghasilkan rekomendasi praktis bagi PT XYZ menyusun perencanaan produksi, distribusi dan strategi keputusan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Berikut merupakan gambaran tahapan sistematis penelitian yang akan dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Langkah-Langkah Penelitian

2.2 Identifikasi Masalah

Identifikasi masalah melalui observasi terhadap fenomena industri, serta analisis laporan data historis produksi kelapa sawit. Adanya beberapa tantangan yang dihadapi muncul yaitu terjadinya fluktuasi hasil produksi panen, tidak adanya sistem prediksi yang optimal, dan sulit untuk melakukan perencanaan strategi selanjutnya. Masalah ini sudah dibahas dengan rinci pada pendahuluan, sehingga bagian ini hanya akan menegaskan bahwa permasalahan dilakukan dengan pendekatan literatur dan sekunder.

2.3 Studi Literatur

Tahapan ini mengumpulkan referensi dan kajian teori yang berkaitan dengan permasalahan penelitian yang dilakukan. Tahapan ini melakukan review sumber ilmiah seperti jurnal, artikel penelitian, buku dan penelitian terdahulu yang berhubungan dengan prediksi produksi kelapa sawit, serta analisis pengaruh pupuk terhadap hasil produksi. Studi literatur digunakan untuk memahami konsep dasar, metode hingga pendekatan yang dilakukan peneliti terdahulu. Sehingga penelitian ini dapat membandingkan performa model yang paling bagus.

2.4 Pengumpulan Data

Pengumpulan data penelitian ini yaitu data kuantitatif berupa data historis produksi kelapa sawit yang berdifat time series. Data ini dikumpulkan untuk melatih model dengan pendekatan random forest dan XGBoost, serta melihat korelasi hubungan variabel terhadap hasil produksi dengan SHAP (SHapley Additive exPlanation). Data yang dikumpulkan yaitu Januari 2020 hingga Desember 2025 dengan total 588 data. Rentang waktu 6 tahun diharapkan dapat memberikan representasi terhadap pola produksi.

2.5 Analisa Data

Data dianalisa untuk menentukan informasi pada dataset yang ada. Analisa menggunakan model random forest, XGBoost dan SHAP untuk melihat proses data dengan metode tersebut. Tahapan ini juga mengolah data historis produksi kelapa sawit yang digunakan untuk proses prediksi dan analisis pengaruh jenis pupuk terhadap hasil produksi. Analisis diharapkan dapat menghasilkan model prediksi yang akurat.

2.6 Preprocessing Data

Tahapan preprocessing data yaitu melakukan persiapan data sebelum digunakan dalam proses pemodelan random forest dan XGBoost. Data yang akan digunakan yaitu data multivariate time series yang terdiri dari variabel tahun. Bulan, umur tanaman, tahun tanam, luas lahan, jumlah pokok produktif, serta data produksi. Tahapan ini akan meliputi yaitu data selection, data cleaning, transformasi data dan feature engineering. Berikut merupakan data Januari hingga Februari 2025 pada Tabel 1.

Tabel 1. Data Januari hingga Februari 2025

Tahun	Bulan	Umur Tanaman	Tahun Tanam	Luas	Pokok Produktif	Produksi
2025	1	20	2005	1305,25	178275	2929700
2025	1	19	2006	3986,33	533479	8848156
2025	1	18	2007	2531,47	336774	5326103
2025	1	17	2008	1555,47	205718	3474811

Tahun	Bulan	Umur Tanaman	Tahun Tanam	Luas	Pokok Produktif	Produksi
2025	1	16	2009	877,74	114166	2171942
2025	1	15	2010	115,56	15501	333858,4
2025	1	12	2013	373,26	50066	845211,4
2025	1	8	2017	356,12	46518	592628,6
2025	1	3	2022	386,07	56741	46390
2025	2	20	2005	1305,25	178275	2650033
2025	2	19	2006	3986,33	533479	8663402
2025	2	18	2007	2531,47	336774	4701483
2025	2	17	2008	1555,47	205718	3125646
2025	2	16	2009	877,74	114166	1786124
2025	2	15	2010	115,56	15501	239881,2
2025	2	12	2013	373,26	50066	816211,8
2025	2	8	2017	356,12	46518	688638,2
2025	2	3	2022	386,07	56741	61000

2.7 Data Selection

Data selection merupakan proses dilakukan pemilihan data yang akan digunakan dalam penelitian ini. Tahapan ini akan dipilih variabel yang berkaitan dengan produksi kelapa sawit, seperti data produksi, tahun dan bulan, umur tanaman, tahun tanam, serta luas pokok produksi.

2.8 Data Cleaning

Tahapan ini yaitu melakukan pembersihan data dari kesalahan seperti missing value, maupun data yang tidak konsisten. Data cleaning dilakukan pengecekan terhadap kualitas data sehingga data yang akan digunakan menjadi lebih akurat dan siap digunakan untuk proses selanjutnya. Penanganan missing value dilakukan dengan pendekatan penghapusan baris untuk setiap data yang missing. Kemudian dilakukan pemeriksaan terhadap duplikasi baris data dan inkonsisten format penulisan nilai numerik. Proses normalisasi atau scaling tidak diterapkan pada data numerik, karena random forest dan XGBoost merupakan model berbasis decision tree yang membentuk aturan split berdasarkan threshold.

2.9 Transformasi Data

Tahapan selanjutnya yaitu transformasi data dengan menggabungkan variabel tahun dan bulan menjadi indeks waktu sehingga data yang diolah dapat digunakan sebagai deret waktu atau time series. Penelitian ini menggunakan data multivariate time series, maka langkah selanjutnya menggunakan teknik sliding window.

2.10 Feature Engineering

Penelitian ini menggunakan data multivariate time series sehingga dibutuhkan sliding window untuk mengubah data time series menjadi format yang sesuai dengan supervised learning yang dimanfaatkan model random forest dan XGBoost. Model random forest dan XGBoost tidak memiliki kemampuan bawaan untuk memahami data time series, sehingga dibutuhkan feature engineering untuk memberikan historis ke dalam setiap baris data. Feature engineering mengambil nilai-nilai beberapa periode waktu sebelumnya (window) dan menjadikan kolom baru pada baris data saat ini untuk model dapat melihat pola historis meskipun diproses data baris per baris secara independen. Penelitian ini akan menggunakan dua sliding window size (sequence length) yaitu 3 bulan dan 6 bulan. Pemilihan sequence dilakukan untuk mengetahui konfigurasi historis data yang paling optimal dalam memprediksi produksi kelapa sawit. Secara matematis, sequence length k ($k=3$ atau $k=6$), pada waktu t dibentuk vektor fitur input X_t dan target output y_t yaitu :

$$X_t = [bulan_t, umur_t, tahun\ tanam_t, luas_t, pokok_t, y_{t-1} \dots, y_{t-k}] \quad (1)$$

2.11. Data Splitting

Penelitian ini membagi dataset menjadi dua bagian yaitu data training dan data testing. Data tersebut digunakan untuk melatih masing-masing model dan mengevaluasi performa pada data yang belum pernah dilihat oleh model sebelumnya. Pada penelitian ini pembagian data dilakukan berdasarkan urutan waktu dengan pembagian 80% untuk data training dan 20% untuk data testing. Seluruh dataset diurutkan terlebih dahulu, kemudian dilakukan pembagian 80:20 dari total rentang waktu tersebut. Pendekatan ini wajib digunakan pada dataset time series karena model hanya menggunakan informasi masa lalu dan menguji masa depan dengan data testing, sehingga menjegah terjadinya kebocoran informasi data.

2.12. Pengujian Random Forest dan XGBoost

Random forest merupakan metode machine learning yang termasuk ke dalam supervised dalam pembentukan kelas pengklasifikasi. Metode ini menggabungkan hasil prediksi dari beberapa pohon keputusan. Setelah melewati proses

tahapan preprocessing produksi kelapa sawit akan diprediksi menggunakan random forest dengan langkah-langkah berikut :

- a. Persamaan impurity node untuk proses split pembentukan decision tree

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right}) \quad (2)$$

Keterangan rumus tersebut dapat diartikan bahwa x_i merupakan variabel yang digunakan untuk split data berdasarkan nilai titik acuan v_{ij} . Pemisahan dari jumlah total data pada node N_s menghasilkan jumlah data pada node kiri n_{left} dan jumlah data pada node kanan n_{right} , di mana tingkatan ketidakmurnian tiap node diukur menggunakan fungsi impurity $H(X)$.

- b. Perhitungan Mean Squared Error (MSE)

$$H(X_{m_i}) = \frac{1}{N_m} \sum_{i \in N_m} (y - \hat{y}_m)^2 \quad (3)$$

Keterangan rumus tersebut menjabarkan bahwa nilai MSE dihitung dari rata-rata kuadrat selisih antara nilai aktual y dan nilai prediksi $\hat{y}_m$ berdasarkan jumlah total data pada node N_m .

- c. Selain MSE dapat menggunakan MAE (Mean Absolute Error)

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y - \hat{y}_m| \quad (4)$$

- d. Setelah proses split, kualitas segmentasi dihitung dengan persamaan berikut

$$G(x, v) = \frac{1}{N_s} (\sum_{y_i \in X_{left}} (y_i - \bar{y}_{left})^2 + \sum_{y_i \in X_{right}} (y_i - \bar{y}_{right})^2) \quad (5)$$

Keterangan rumus sebagai berikut nilai impurity hasil split $G(x, v)$ dihitung dari penjumlahan kuadrat selisih antara nilai aktual node kiri y_i terhadap rata-rata node kiri $\bar{y}_{left}$ serta selisih antara nilai aktual node kanan y_j terhadap rata-rata node kanan $\bar{y}_{right}$.

- e. Hasil prediksi diperoleh prediksi akhir dengan menghitung rata-rata hasil prediksi seluruh decision tree

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K T_k(x) \quad (6)$$

Keterangan rumus tersebut menjabarkan bahwa hasil prediksi akhir $\hat{y}$ diperoleh dari rata-rata hasil prediksi tree ke- k $T_k(x)$ untuk seluruh jumlah decision tree K yang digunakan.

- f. Menghitung tingkat kepentingan variabel

$$n_k = w_k G_k - w_{left} G_{left} - w_{right} G_{right} \quad (7)$$

Keterangan rumus tersebut menjabarkan bahwa nilai tingkat kepentingan node n_k diperoleh dari perkalian proporsi data node utama w_k dengan impurity node utama G_k , dikurangi perkalian proporsi data node kiri w_{left} dan impurity node kiri G_{left} , serta perkalian proporsi data node kanan w_{right} dan impurity node kanan G_{right} .

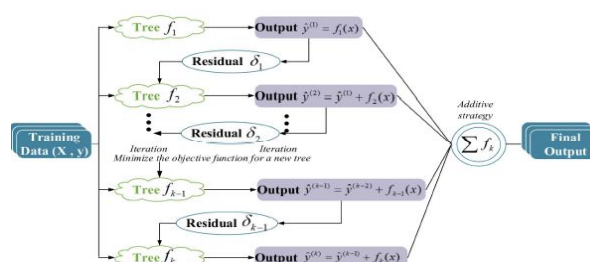
- g. Importance setiap fitur

$$f_i = \frac{\sum_{j \in \text{feature}} \text{in}^j}{\sum_{k \in \text{all nodes}} \text{in}^k} \quad (8)$$

Kemudian dilakukan normalisasi feature importance

$$f_{ni} = \frac{f_i}{\sum_j f_j} \quad (9)$$

Xgboost merupakan pengembangan dari algoritma gradient boosting versi yang lebih canggih dan cepat. Metode ini menggabungkan decision tree secara bertahap dimana pohon dibangun untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Setiap iterasi XGBoost menghitung residual error dari model sebelumnya dan melakukan perbaikan pada iterasi berikutnya. XGBoost menerapkan teknik regularization dan pruning untuk meningkatkan kualitas model. Metode ini perlu pengoptimalan parameter model dengan meminimalkan nilai loss function pada setiap iterasi. Gambar 2 berikut merupakan arsitektur dari XGBoost.



Gambar 2. Arsitektur XGBoost [10]

2.13. SHAP

SHAP (SHapley Additive exPlanation) merupakan metode interpretabilitas model yang dikembangkan berdasarkan konsep shapley value dalam teori (game theory). Metode ini menjelaskan kontribusi masing-masing fitur terhadap hasil prediksi yang dihasilkan machine learning. Secara matematis nilai shapley dihitung mempretimbangkan perubahan nilai prediksi ketika suatu fitur ditambahkan ke subset fitur lainnya, semakin besar nilai SHAP maka semakin besar pula nilai pengaruh fitur tersebut begitupun sebaliknya. Berikut rumus sistematis dari SHAP :

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [V(S \cup \{j\}) - V(S)] \quad (10)$$

Keterangan rumus tersebut menjabarkan bahwa nilai SHAP untuk fitur ke- j ϕ_j dihitung dari penjumlahan kontribusi marginal fitur j , yaitu selisih antara nilai prediksi saat fitur dalam subset S beserta fitur j digunakan $V(S \cup \{j\})$ dengan nilai prediksi saat hanya fitur dalam subset S yang digunakan $V(S)$. Perhitungan ini disesuaikan dengan bobot kombinasi dari subset fitur yang tidak mengandung fitur j (S) terhadap seluruh fitur dalam model (N).

Dari sifat efficiency, diperoleh:

$$f^{(i)} = \phi_0 + \sum_j = 1M\phi_{i,j} \quad (11)$$

Keterangan rumus tersebut menjabarkan bahwa prediksi model untuk observasi ke- i $f^{(i)}$ diperoleh dari penjumlahan antara rata-rata prediksi model ϕ_0 dengan total kontribusi fitur ke- j pada observasi ke- i $\phi_{i,j}$ untuk seluruh jumlah fitur M .

2.14. Evaluasi

Pengujian penelitian ini berfungsi untuk melihat seberapa jauh model yang dibangun mampu memprediksi data produksi dan melihat hubungan faktor jenis pupuk dengan produksi dilakukan baik. MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), dan R2 (Coefficient of Determination) untuk evaluasi prediksi kelapa sawit. Berikut rumus untuk mengukur perhitungan :

- a. MAE (Mean Absolute Error)

Mae adalah mengukur rata-rata besar kesalahan prediksi tanpa memperhatikan arah kesalahan.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

Keterangan rumus tersebut menjabarkan bahwa nilai MAE dihitung dari rata-rata selisih absolut antara nilai aktual y_i dan nilai prediksi $\hat{y}_i$ berdasarkan jumlah total data n . Semakin kecil nilai MAE, maka performa model dianggap semakin baik karena rata-rata tingkat kesalahannya semakin rendah.

- b. RMSE (Root Mean Squared Error)

RMSE adalah mengukur besar error model dengan memberikan pinalti lebih besar pada error yang besar.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (13)$$

Keterangan rumus tersebut menjabarkan bahwa nilai RMSE dihitung dari akar rata-rata kuadrat error, yaitu selisih antara nilai aktual y_i dan nilai prediksi $\hat{y}_i$ yang dikuadratkan untuk seluruh jumlah data n . Semakin kecil nilai RMSE, semakin akurat model yang dihasilkan. Nilai ini sangat sensitive terhadap error berukuran besar karena adanya proses pengkuadratan.

- c. R2 (Coefficient of Determination)

R2 digunakan untuk mengukur seberapa baik model menjelaskan variasi data aktual.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

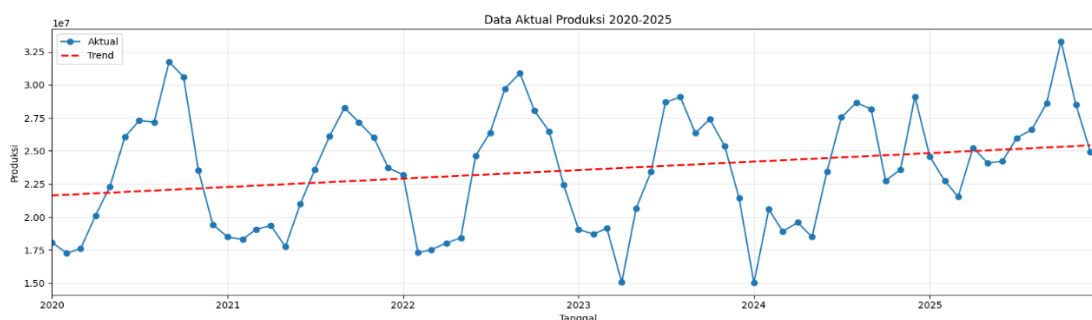
Keterangan rumus tersebut menjabarkan bahwa nilai R^2 dihitung dari 1 dikurangi rasio antara total kuadrat selisih nilai aktual y_i terhadap nilai prediksi $\hat{y}_i$ dengan total kuadrat selisih nilai aktual y_i terhadap rata-rata nilai aktual $\bar{y}$. Nilai ini digunakan untuk mengukur seberapa baik model dalam menjelaskan variasi dari data aktual. Semakin mendati nilai 1, semakin baik performa model.

3. HASIL DAN PEMBAHASAN

Data produksi kelapa sawit yang telah didapatkan berasal dari perusahaan PT XYZ yang berada di Riau. Data tersebut berjumlah 588 data dari bulan Januari 2020 hingga Desember 2025, sehingga data 6 tahun terakhir produksi kelapa sawit dapat membantu proses penelitian ini. Berikut Tabel 2, merupakan data Januari 2020 hingga Desember 2025. Tabel 2 merupakan sampel data produksi kelapa sawit pada 6 tahun terkahir yaitu Januari 2020 hingga Desember 2025. Data tersebut akan menjadi data training dan data testing pada metode random forest dan XGBoost. Data pada tabel 2 akan diolah untuk mendapatkan grafik data aktual.

Tabel 2. Sampel data Januari 2020 hingga Desember 2025

Tahun	Bulan	Umur Tanaman	Tahun Tanam	Luas	Pokok Produktif	Produksi
2020	1	15	2005	1305,25	178275	2242685
2020	1	14	2006	3986,33	529494	6788537
2020	1	13	2007	2983,98	392416	4137907
2020	1	12	2008	2298,43	302114	2681955
2020	1	11	2009	984,47	127847	1267843
2020	1	10	2010	115,56	15471	197064,2
2020	1	7	2013	373,26	49431	659030
2020	1	3	2017	356,12	46518	112770
--	--	--	--	--	--	--
2025	12	3	2022	386,07	56741	280044

**Gambar 3.** Plot Grafik Data Aktual Produksi Kelapa Sawit

Pada Gambar 3 merupakan grafik data aktual produksi kelapa sawit Januari 2020 hingga Desember 2025 yang telah dilakukan penyeleksian data yang dipilih. Kemudian data tersebut akan digunakan metode random forest dan XGBoost untuk perbandingan pengujian metode, serta melakukan analisis korelasi menggunakan SHAP. Pada gambar diatas sumbu horizontal menunjukan periode sedangkan sumbu vertical untuk hasil produksi kelapa sawit. Terdapat trend pada gambar tersebut yang menunjukan adanya indikasi kenaikan dan peningkatan produksi kelapa sawit dari tahun 2020 hingga 2025.

3.1 Hasil Penelitian

3.1.1 Feature Engineering

Random forest dan XGBoost tidak memiliki kemampuan bawaan untuk memahami data time series, sehingga dibutuhkan feature engineering untuk memberikan historis ke dalam setiap baris data. Sebelum diterapkan feature engineering, dataset produksi kelapa sawit memiliki kolom-kolom dasar seperti tahun, bulan, tahun tanam, serta produksi. Berikut data sebelum feature engineering.

Tabel 3. Dataset sebelum feature engineering

Date idx	Tahun	Bulan	Tahun Tanam	Produksi
2020-01-01	2020	1	2005	2,392,032.31
2020-02-01	2020	2	2005	2,720,847.73
2020-03-01	2020	3	2005	3,099,512.05
2020-04-01	2020	4	2005	2,972,897.04

Tabel 3 menunjukan dataset sebelum diterapkannya feature engineering. Kemudian setelah diterapkan teknik sliding window dengan size 3, terbentuk 5 kolom baru, yaitu lag_1, Lag_2, Lag_3 yang merepresentasikan nilai produksi 1 hingga 3 bulan sebelumnya, serta roll_mean_3 dan roll_std_3 yang masing-masing merepresentasikan rata-rata dan variasi produksi selama 3 bulan terakhir. Berikut merupakan contoh hasil yang ditampilkan pada tabel 4.

Tabel 4. Dataset sesudah feature engineering (window size = 3)

Date idx	Produksi	Lag 1	Lag 2	Lag 3	Roll mean 3	Roll std 3
2020-01-01	2,392,032.31	2,300,663.49	2,413,774.84	2,242,684.84	2,319,041	87,012.9
2020-02-01	2,720,847.73	2,392,032.31	2,300,663.49	2,413,774.84	2,368,824	60,021.1
2020-03-01	3,099,512.05	2,720,847.73	2,392,032.31	2,300,663.49	2,471,181	220,991.2
2020-04-01	2,972,897.04	3,099,512.05	2,720,847.73	2,392,032.31	2,737,464	354,032.4

Tabel 4 terlihat bahwa baris data untuk bulan Januari hingga Maret 2020 tidak muncul pada tabel karena belum memiliki data historis 3 bulan yang lengkap sehingga dibuang pada tahap pembersihan data. Perlakuan yang sama juga diterapkan dengan window size 6 menghasilkan 8 kolom baru yaitu lag_1 hingga lag_6, roll_mean_6 dan

roll_std_6 yang mana semakin besar window size yang digunakan maka semakin banyak pula jumlah baris data awal yang harus dibuang karena membutuhkan historis yang lebih panjang. Berikut Tabel 5a dan 5b dataset sesudah feature engineering (window size = 6).

Tabel 5a. Dataset sesudah feature engineering (window size = 6)

Date_idx	Produksi	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6
2020-07-01	2,972,897.04	3,099,512.05	2,720,847.73	2,392,032.31	2,300,663.49	2,413,774.84	2,242,684.84
2020-08-01	3,288,608.73	2,972,897.04	3,099,512.05	2,720,847.73	2,392,032.31	2,300,663.49	2,413,774.84
2020-09-01	3,985,790.37	3,288,608.73	2,972,897.04	3,099,512.05	2,720,847.73	2,392,032.31	2,300,663.49

Tabel 5b. Lanjutan Dataset sesudah feature engineering (window size = 6)

Roll_mean_6	Roll_std_6
2,528,253	325,096.0
2,649,955	333,385.2
2,795,760	395,037.0

Dengan diterapkannya feature engineering ini, model random forest dan xgboost secara alami tidak memiliki kemampuan memahami urutan waktu, kini dapat mempertimbangkan informasi historis produksi pada setiap baris data, sehingga lebih sesuai untuk digunakan pada data time series.

3.1.2 Random Forest

Random forest merupakan metode ensemble learning berbasis decision tree yang bekerja dengan membangun sejumlah pohon keputusan secara acak dan mengabungkan hasil prediksi dari seluruh pohon untuk menghasilkan prediksi akhir yang lebih akurat dan stabil [14]. Penelitian ini menggunakan proses hyperparameter tuning terhadap parameter estimators, max depth dan samples split. Adapun konfigurasi hyperparameter yang digunakan dapat dilihat pada tabel 6 berikut ini.

Tabel 6. Konfigurasi Hyperparameter Random Forest

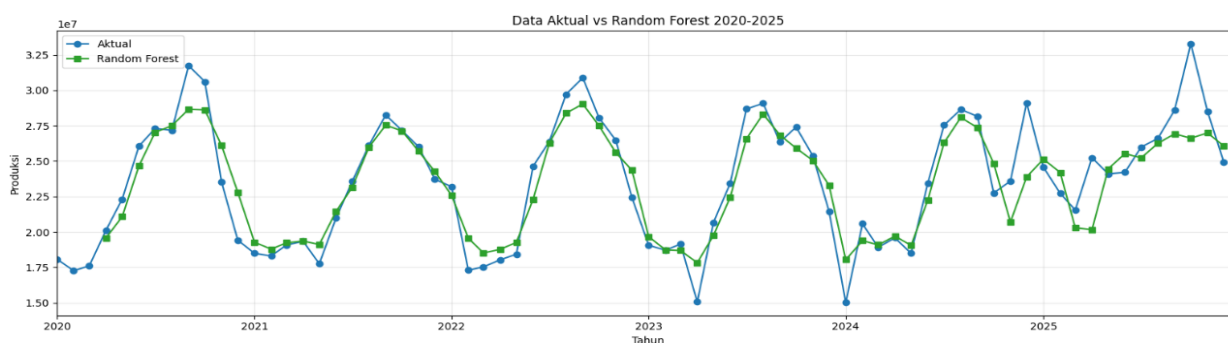
Hyperparameter	Nilai yang Diuji
n_estimators	100, 200
max_depth	5, 10, None
min_samples_split	2, 5

Tabel 6 merupakan konfigurasi proses tuning hyperparameter, kemudian melakukan pengujian sequence length yaitu 3 dan 6 bulan. Hasil evaluasi model random forest ditunjukkan pada tabel 7 berikut.

Tabel 7. Hasil Evaluasi Random Forest Berdasarkan Sequence Length Data Training

Sequence Length	MAE	RMSE	R ²
3	373.758	644.373	0,949
6	444.522	721.270	0,937

Berdasarkan Tabel 7 dapat dilihat bahwa sequence length 3 menghasilkan performa terbaik dengan nilai MAE sebesar 373.758, RMSE sebesar 644.373 dan nilai R² sebesar 0,949 pada data training. Nilai RMSE yang lebih kecil menunjukkan bahwa kesalahan prediksi model relatif rendah, sedangkan nilai R² sebesar 0,949 menunjukkan bahwa model mampu menjelaskan sekitar 95% variasi data produksi kelapa sawit. Dengan demikian, sequence length 3 dipilih sebagai konfigurasi terbaik pada model Random Forest untuk digunakan pada tahap evaluasi dan forecasting. Hasil perbandingan antara data aktual dan hasil prediksi Random Forest ditunjukkan pada Gambar 4.



Gambar 4. Perbandingan Data Aktual dan Prediksi Random Forest Tahun 2020–2025

Pada Gambar 4 terlihat bahwa model Random Forest mampu mengikuti pola umum pergerakan produksi kelapa sawit. Prediksi yang dihasilkan memiliki pola yang relatif serupa dengan data aktual, terutama pada periode

yang menunjukkan pola musiman tahunan. Namun demikian, pada beberapa titik puncak dan lembah produksi terlihat bahwa model menghasilkan prediksi yang lebih halus (smoothing) dibandingkan data aktual. Hal ini menunjukkan bahwa Random Forest memiliki kecenderungan mengurangi fluktuasi ekstrem sehingga beberapa lonjakan produksi tidak sepenuhnya dapat ditangkap oleh model [15].

3.1.3 XGBoost

XGBoost merupakan algoritma ensemble berbasis gradient boosting yang membangun pohon keputusan secara bertahap untuk memperbaiki kesalahan prediksi yang dihasilkan oleh pohon sebelumnya [16]. Pendekatan ini memungkinkan model mempelajari hubungan nonlinier yang kompleks pada data time series. Pada penelitian ini dilakukan proses tuning terhadap parameter `n_estimators`, `max_depth`, `learning_rate`, `gamma`, dan `reg_lambda`. Konfigurasi hyperparameter yang diuji dapat dilihat pada Tabel 8.

Tabel 8. Konfigurasi Hyperparameter XGBoost

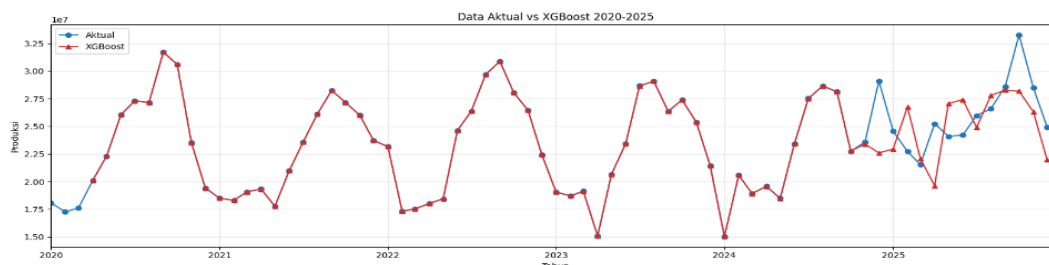
Hyperparameter	Nilai yang Diuji
<code>n_estimators</code>	50, 100, 200
<code>max_depth</code>	3, 5, 6, 7, 8
<code>learning_rate</code>	0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9
<code>gamma</code>	0, 0.2, 0.4, 0.6, 0.8, 1
<code>reg_lambda</code>	0, 0.2, 0.4, 0.6, 0.8, 1

Hasil evaluasi model XGBoost berdasarkan sequence length ditunjukkan pada Tabel 9

Tabel 9. Hasil Evaluasi XGBoost Berdasarkan Sequence Length Data Training

Sequence Length	MAE	RMSE	R ²
3	543.344	857.743	0,911
6	642.664	1.006.565	0,877

Berdasarkan Tabel 9 dapat dilihat bahwa sequence length 3 menghasilkan performa terbaik dibandingkan sequence length lainnya dengan nilai RMSE sebesar 857.743 dan nilai R² sebesar 0,911 pada data training. Hasil tuning menunjukkan bahwa kombinasi hyperparameter terbaik diperoleh pada sequence length 3 dengan parameter `gamma` = 0 dan `learning_rate` sebesar 0,7 serta kombinasi parameter lainnya yang menghasilkan performa prediksi paling optimal pada data pengujian. Visualisasi hasil prediksi model XGBoost dibandingkan data aktual ditunjukkan pada Gambar 5.



Gambar 5. Perbandingan Data Aktual dan Prediksi XGBoost Tahun 2020–2025

Berdasarkan Gambar 5 terlihat bahwa model XGBoost mampu mengikuti pola aktual dengan sangat baik. Pada sebagian besar periode, garis prediksi hampir berimpit dengan garis aktual sehingga menunjukkan kemampuan model dalam menangkap pola historis produksi kelapa sawit. Kemampuan ini menunjukkan bahwa mekanisme boosting yang dimiliki XGBoost mampu memperbaiki kesalahan prediksi secara bertahap sehingga menghasilkan akurasi yang tinggi [17].

3.1.4 Perbandingan Kinerja Model Data Testing

Setelah diperoleh model terbaik dari masing-masing algoritma, dilakukan evaluasi menggunakan data testing untuk membandingkan performa kedua model.

Tabel 10. Perbandingan Kinerja Model

Model	MAE	RMSE	R ²
Random Forest	1.283.547	1.792.539	0,824
XGBoost	553.293	1.496.990	0,877

Berdasarkan Tabel 10 model XGBoost menghasilkan performa yang lebih baik dibandingkan Random Forest. Hal ini ditunjukkan oleh nilai MAE dan RMSE yang lebih rendah serta nilai koefisien determinasi yang lebih tinggi. Nilai RMSE XGBoost sebesar 1.496.990 lebih rendah dibandingkan Random Forest sebesar 1.792.539. Selain itu,

nilai R^2 XGBoost mencapai 0,877 sedangkan Random Forest hanya sebesar 0,824. Hasil ini menunjukkan bahwa XGBoost mampu menjelaskan variasi data produksi dengan lebih baik dibandingkan Random Forest. Keunggulan XGBoost disebabkan oleh mekanisme gradient boosting yang memungkinkan model memperbaiki kesalahan prediksi pada setiap iterasi sehingga lebih efektif dalam menangkap pola nonlinier dan fluktuasi produksi kelapa sawit [18]. Berdasarkan hasil evaluasi tersebut, model XGBoost dipilih sebagai model terbaik dalam penelitian ini.

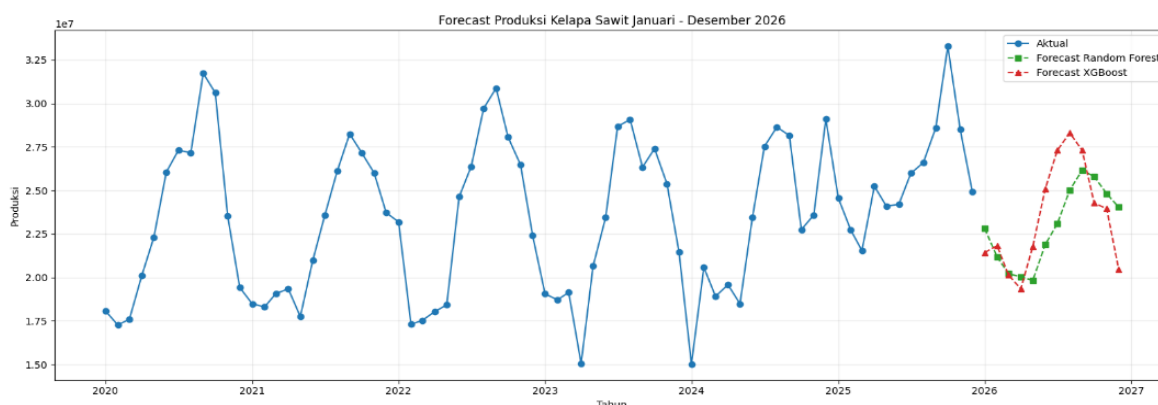
3.1.5 Forecast Produksi Kelapa Sawit

Setelah model terbaik diperoleh, dilakukan proses forecasting untuk memprediksi total produksi kelapa sawit pada periode Januari hingga Desember 2026. Forecast dilakukan menggunakan model Random Forest dan XGBoost terbaik yang diperoleh dari tahap sebelumnya. Hasil forecasting ditunjukkan pada Tabel 11.

Tabel 11. Hasil Forecast Produksi

Model	Periode	Forecast Produksi
RF	2026-01	2.280991e+07
RF	2026-02	2.117914e+07
RF	2026-03	2.022512e+07
RF	2026-04	2.002874e+07
RF	2026-05	1.983451e+07
RF	2026-06	2.189483e+07
RF	2026-07	2.307909e+07
RF	2026-08	2.502092e+07
RF	2026-09	2.616246e+07
RF	2026-10	2.579806e+07
RF	2026-11	2.481779e+07
RF	2026-12	2.404113e+07
XGboost	2026-01	2.142949e+07
XGboost	2026-02	2.182859e+07
XGboost	2026-03	2.014603e+07
XGboost	2026-04	1.934711e+07
XGboost	2026-05	2.179260e+07
XGboost	2026-06	2.509531e+07
XGboost	2026-07	2.730149e+07
XGboost	2026-08	2.831437e+07
XGboost	2026-09	2.730078e+07
XGboost	2026-10	2.427337e+07
XGboost	2026-11	2.395934e+07
XGboost	2026-12	2.047652e+07

Visualisasi hasil forecasting ditunjukkan pada Gambar 6.



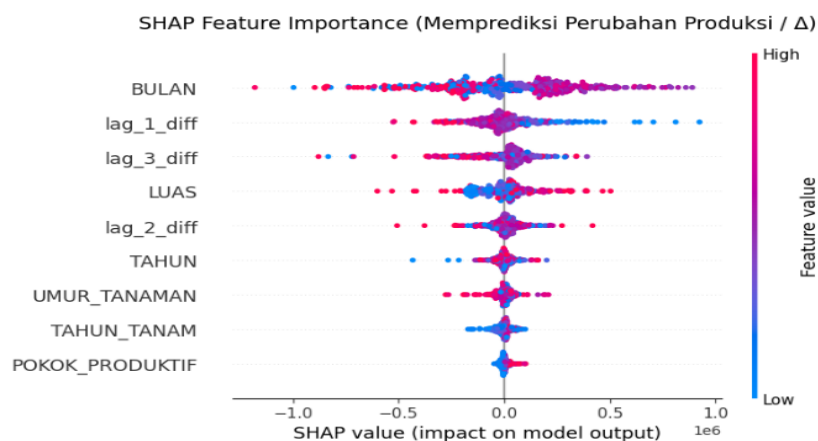
Gambar 6. Forecast Produksi Kelapa Sawit Januari hingga Desember 2026

Berdasarkan hasil forecasting, model Random Forest memprediksi total produksi kelapa sawit pada Januari 2026 sebesar 2.280991e+07 kg, sedangkan model XGBoost memprediksi sebesar 2.142949e+07 kg. Perbedaan hasil prediksi tersebut menunjukkan bahwa masing-masing model memiliki pendekatan yang berbeda dalam mempelajari pola historis data. Random Forest cenderung menghasilkan prediksi yang lebih tinggi karena melakukan agregasi dari banyak pohon keputusan yang mempertimbangkan berbagai kemungkinan pola data [19]. Sebaliknya, XGBoost menghasilkan prediksi yang lebih konservatif karena proses boosting yang lebih fokus pada pengurangan kesalahan prediksi pada data pelatihan. Meskipun nilai forecast yang dihasilkan berbeda, kedua model masih menunjukkan

bahwa produksi kelapa sawit pada Januari 2026 diperkirakan berada pada kisaran 19–23 juta kilogram. Mengingat hasil evaluasi sebelumnya menunjukkan bahwa XGBoost memiliki nilai RMSE yang lebih rendah dan nilai R^2 yang lebih tinggi, maka hasil forecasting dari model XGBoost dianggap lebih representatif untuk digunakan sebagai acuan prediksi produksi periode Januari hingga Desember 2026.

3.1.6 Analisis Pengaruh Variabel Menggunakan SHAP

SHAP (SHapley Additive exPlanations) digunakan untuk menginterpretasikan hasil prediksi model XGBoost dan mengidentifikasi kontribusi masing-masing variabel terhadap produksi kelapa sawit. Metode SHAP menghitung kontribusi setiap fitur terhadap nilai prediksi yang dihasilkan model sehingga dapat diketahui faktor yang paling berpengaruh terhadap produksi [20]. Hasil analisis SHAP Summary Plot ditunjukkan pada Gambar 7.

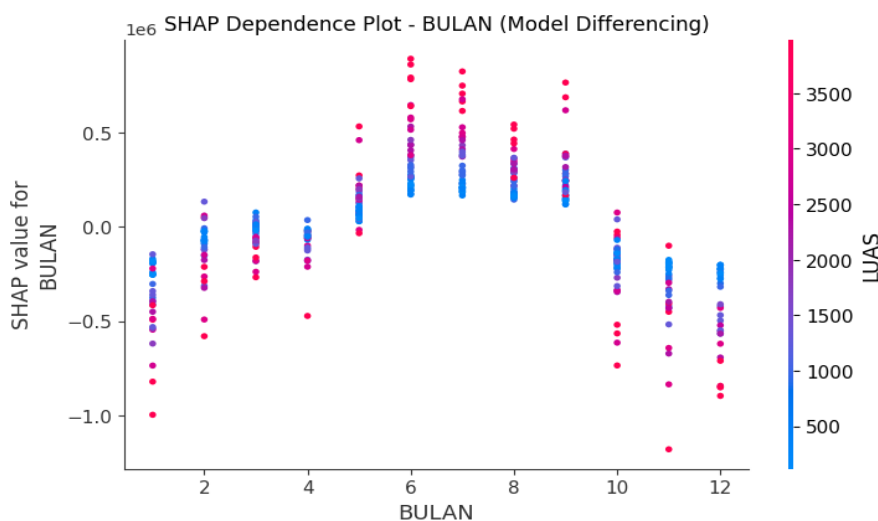


Gambar 7. SHAP Summary Plot Faktor Variabel terhadap Produksi Kelapa Sawit

Berdasarkan Gambar 7, terlihat bahwa **Bulan** memiliki pengaruh terbesar terhadap prediksi produksi kelapa sawit. Hal ini ditunjukkan oleh penyebaran nilai SHAP yang paling luas dibandingkan fitur lainnya. Selanjutnya diikuti oleh lag_1, lag_3 dan luas. Sementara itu, las_2, tahun, umur tanaman, tahun tanam dan pokok produktif menunjukkan pengaruh yang relatif lebih kecil karena memiliki rentang nilai SHAP yang lebih sempit. Selain menunjukkan tingkat kepentingan fitur, SHAP Summary Plot juga memperlihatkan arah pengaruh masing-masing variabel. Titik berwarna merah menunjukkan nilai fitur yang tinggi, sedangkan titik berwarna biru menunjukkan nilai fitur yang rendah. Sebagian besar titik merah pada fitur bulan berada di sisi kanan grafik (nilai SHAP positif), yang mengindikasikan bahwa bulan cenderung meningkatkan prediksi produksi kelapa sawit.

3.1.7 Analisis Pengaruh Bulan terhadap Produksi

Dapat memahami hubungan antara penggunaan bulan dan produksi kelapa sawit secara lebih rinci, dilakukan analisis menggunakan SHAP Dependence Plot sebagaimana ditunjukkan pada Gambar 8.



Gambar 8. SHAP Dependence Plot Bulan

Berdasarkan Gambar 8, terlihat adanya hubungan positif yang kuat antara bulan dan nilai SHAP. Dependence plot menunjukkan pola musiman yang jelas pada variabel bulan dimana produksi cenderung menurun pada awal tahun (Januari-April), meningkat signifikan dan mencapai puncaknya pada pertengahan tahun (Mei-September, khususnya

Juni-Juli). Kemudian menurun kembali pada akhir tahun (Oktober-Desember). Pola ini konsisten di berbagai ukuran luas kebun, menunjukkan bahwa faktor musiman berpengaruh secara merata terlepas dari skala kebun. Temuan ini sejalan dengan karakteristik siklus panen kelapa sawit yang umumnya mencapai puncak produksi pada pertengahan tahun.

3.1.8 Eksperimen Tambahan Variabel Jenis dan Dosis Pupuk

Dapat Selain data utama yang digunakan variabel tahun, bulan, umur tanaman, tahun tanam, luas dan pokok produktif. Penelitian ini juga menguji pendekatan alternatif dengan menambahkan variabel jenis dan dosis pemupukan sebagai fitur yaitu urea, rp peru, mop, kieserite, hgfb, dolomite, CuSO₄, ZnSO₄, dan NPK. Berikut data pupuk dari Januari 2020 hingga Desember 2025 untuk analisis korelasi produksi terhadap jenis pupuk.

Tabel 12. Sampel data Januari 2020 hingga Desember 2025

Tahun/Bulan	Produksi	Urea	RP Peru	MOP	Kieserite	HGFB	Dolomite	CuSO ₄	ZnSO ₄	NPK
2020/1	2242685	27772	0	0	0	0	0	0	0	0
2020/1	6788537	33709	0	114904	0	0	0	0	0	0
2020/1	4137907	73900	0	97204	0	0	0	0	0	0
2020/1	2681955	27322	0	58496	0	0	477	0	0	0
2020/1	1267843	8890	0	15989	0	0	21302	0	0	0
2020/1	197064,2	0	0	0	0	0	0	0	0	0
2020/1	659030	0	0	0	0	0	0	0	0	0
2020/1	112770	0	0	0	0	0	0	0	0	0
--	--	--	--	--	--	--	--	--	--	--
2025/12	280044	0	0	0	0	0	0	0	0	0

Tabel 12 merupakan data sampel produksi dan jenis pupuk dari Januari 2020 hingga Desember 2025. Data tersebut akan digunakan sebagai analisis korelasi produksi terhadap jenis pupuk menggunakan teknik SHAP. Model random forest dan xgboost dilatih menggunakan variabel tersebut dengan hasil evaluasi pada data training sebagai berikut.

Tabel 13. Perbandingan Kinerja Model Pupuk

Model	MAE	RMSE	R ²
Random Forest	1,418,621	2,112,878	0.4857
XGBoost	1,534,058	2,136,976	0.4739

Berdasarkan Tabel 13 hasil tersebut, kedua model hanya mampu menjelaskan sekitar 47%-49% variasi produksi dilihat dari R², dengan tingkat kesalahan prediksi (RMSE) yang relatif tinggi dibanding model utama penelitian ini. Berdasarkan pertimbangan tersebut, penelitian ini selanjutnya difokuskan pada tujuan variabel utama yaitu tahun, bulan, umur tanaman, tahun tanam, luas, pokok produktif hingga produksi. Secara empiris variabel tersebut menghasilkan performa model yang lebih baik dan stabil dibandingkan dengan pendekatan yang menyertakan variabel pupuk.

3.2 Pembahasan

Berdasarkan hasil pengujian yang dilakukan, model XGBoost menunjukkan performa yang lebih baik dibandingkan Random Forest dalam memprediksi produksi kelapa sawit di PT XYZ, dengan nilai R² sebesar 0,877, MAE 553.293, dan RMSE 1.496.990 (sedangkan Random Forest memperoleh R² 0,824, MAE 1.283.547, dan RMSE 1.792.539). Keunggulan XGBoost ini sejalan dengan temuan penelitian terdahulu pada konteks prediksi produksi pertanian yang menemukan bahwa XGBoost memberikan performa terbaik dengan R² 0,89 [11]. Mekanisme *boosting* yang bekerja secara iteratif dalam memperbaiki kesalahan prediksi dinilai lebih adaptif dalam menangkap pola produksi dibandingkan *bagging* pada Random Forest. Kendati demikian, temuan ini juga menegaskan bahwa performa algoritma berbasis *boosting* dan *bagging* sangat fleksibel bergantung pada jenis data sebagaimana pada studi perbandingan CPO terdahulu, XGBoost tidak selalu unggul mutlak atas model lain seperti LSTM sebelum dilakukan *tuning* [10]. Selain itu, penggunaan analisis SHAP berhasil mengungkap bahwa variabel bulan serta fitur *lag* (*lag_1* dan *lag_3*) menjadi faktor paling dominan terhadap hasil prediksi, sekaligus mengatasi keterbatasan studi terdahulu yang umumnya belum mengintegrasikan interpretabilitas model pada pemodelan produksi perkebunan [9], [10], [11], [12]. Kebaharuan penelitian ini terletak pada penggunaan data primer PT XYZ yang dikombinasikan dengan evaluasi performa model dan analisis SHAP secara bersamaan. Berbeda dengan penelitian sebelumnya yang dominan menggunakan data sekunder, pendekatan ini tidak hanya memberikan tingkat akurasi prediksi periode Januari–Desember 2026 yang lebih presisi, tetapi juga menghasilkan wawasan praktis mengenai pola musiman sebagai acuan langsung bagi manajemen PT XYZ dalam menyusun target produksi bulanan, strategi distribusi, dan perencanaan operasional.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, metode Random Forest dan XGBoost berhasil diimplementasikan untuk melakukan forecasting produksi kelapa sawit pada PT XYZ menggunakan data historis produksi, karakteristik tanaman, serta data penggunaan pupuk. Hasil pengujian menunjukkan bahwa kedua model memperoleh performa terbaik pada sequence length 3 bulan. Dari hasil evaluasi, model XGBoost memberikan kinerja yang lebih baik dibandingkan Random Forest dengan nilai MAE sebesar 553.293, RMSE sebesar 1.496.990, dan R^2 sebesar 0,877, sedangkan Random Forest menghasilkan nilai MAE sebesar 1.283.547, RMSE sebesar 1.792.539, dan R^2 sebesar 0,824. Dengan demikian, XGBoost dipilih sebagai model terbaik dalam penelitian ini karena mampu menghasilkan tingkat kesalahan prediksi yang lebih rendah dan tingkat akurasi yang lebih tinggi. Keunggulan ini konsisten dengan karakteristik mekanisme boosting yang memperbaiki kesalahan prediksi secara bertahap sehingga lebih adaptif, sementara random forest kurang menangkap lonjakan produksi ekstream. Berdasarkan model terbaik yang diperoleh, dilakukan forecasting produksi kelapa sawit untuk periode Januari 2026 hingga Desember 2026. Hasil prediksi menunjukkan bahwa model Random Forest menghasilkan estimasi produksi sebesar $2.280991e+07$ kg pada Januari 2026, sedangkan model XGBoost menghasilkan estimasi sebesar $2.142949e+07$ kg pada Januari 2026. Mengingat performa evaluasi XGBoost lebih baik dibandingkan Random Forest, maka hasil prediksi dari model XGBoost dianggap lebih representatif sebagai acuan forecasting produksi pada periode tersebut. Selain melakukan forecasting, penelitian ini juga menerapkan metode SHAP (SHapley Additive exPlanations) untuk menginterpretasikan pengaruh masing-masing variabel terhadap produksi kelapa sawit. Hasil analisis SHAP menunjukkan bahwa bulan merupakan faktor yang memiliki pengaruh paling besar terhadap prediksi produksi, diikuti oleh lag_1, lag_3 dan luas. SHAP Dependence Plot juga menunjukkan adanya hubungan positif antara variabel bulan dan peningkatan produksi. Hasil forecasting ini dapat dimanfaatkan PT XYZ sebagai acuan dalam menyusun target produksi bulanan. Secara keseluruhan, penelitian ini memberikan kontribusi baik dari sisi akurasi prediksi maupun interpretabilitas model dan dapat mendukung perusahaan XYZ dalam pengambilan keputusan berbasis data terkait strategi operasional. Kontribusi penelitian ini terletak pada penyediaan prebandingan random forest dan XGBoost menggunakan data primer produksi kelapa sawit serta penerapan SHAP yang menghasilkan wawasan praktis mengenai pola musiman sebagai dasar perencanaan operasional PT XYZ, sehingga dapat menjadi rujukan bagi penelitian sejenis pada perusahaan Perkebunan lainnya. Penelitian selanjutnya dapat mempertimbangkan penambahan variabel eksternal seperti curah hujan dan kondisi iklim untuk meningkatkan akurasi prediksi pada data produksi kelapa sawit.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: R. A. S., dan D. K. D.;

Metodologi: R. A. S., dan D. K. D.;

Perangkat Lunak: R. A. S.;

Validasi: R. A. S., dan D. K. D.;

Analisis Formal: R. A. S., dan D. K. D.;

Investigasi: R. A. S.;

Sumber Daya: C. K. D.;

Kurasi Data: R. A. S.;

Penulisan Draf Awal: R. A. S.;

Penulisan, Peninjauan, dan Penyuntingan: R. A. S., dan D. K. D.;

Visualisasi: R. A. S.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini dilakukan secara mandiri.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] M. Irpan and U. Rio, "Model Prediksi Jumlah Produksi Kelapa Sawit Menggunakan Regresi Linear Berganda Di PT.Surya Argolika Rekasa," *Jurnal Ilmiah B E T R I K : Besemah Teknologi Informasi dan Komputer*, vol. 16, no. 02, hal. 122–130, Agustus 2025, doi: 10.36050/dv9sd116
- [2] Ismiasih dan H. Afroda, "Faktor Penentu Produksi Kelapa Sawit Rakyat Di Provinsi Riau," *Jurnal Penelitian Pertanian Terapan*, vol. 23, no. 2, hal. 211–218, 2023, doi: 10.25181/jppt.v23i2.2726.
- [3] A. P. Putra and A. M. Harahap, "Implementasi Time Series Forecasting dengan Algoritma LSTM untuk Pemantauan dan Prediksi Produktivitas Kelapa Sawit Berdasarkan Hasil Panen," *JURIKOM (Jurnal Riset Komputer)*, vol. 12, no. 2, pp. 46–56, Apr. 2025, doi: 10.30865/jurikom.v12i2.8495.
- [4] T. Susanti and R. Masdalipa, "Estimasi Produksi Kelapa Sawit PT. Arta Prigel Lahat Menggunakan Algoritma Deep Learning," *Digital Transformation Technology (Digitech)*, vol. 6, no. 1, p. 73, Feb. 2026, doi: 10.47709/digitech.v6i1.7807.
- [5] W. D. U. Parwati, F. H. Nadeak, and V. Kautsar, "Analisis Pertumbuhan dan Produktivitas Kelapa Sawit pada Variasi Kerapatan Tanam," *Jurnal Agro Industri Perkebunan*, vol. 12, no. 2, hal. 105–116, Jul. 2023, doi: 10.25181/jaip.v12i2.3535.
- [6] D. Desrihastuti, A. Maryanti, T. E. Sabli, I. A. Mahendra, dan N. A. Hardi, "Dampak Kemiringan Lahan Terhadap Kadar Hara dan Produksi Kelapa Sawit," *Jurnal Agro Industri Perkebunan*, vol. 12, no. 2, hal. 59–70, 2024, doi: 10.25181/jaip.v12i2.3454.
- [7] Rossy Mirasari, P. Puspita, and R. Ramli, "Analisis Korelasi Terhadap Produksi Kelapa Sawit (*Elaeis guineensis* Jacq.) Dan Pemupukan NPK," *Jurnal Agrisistem*, vol. 19, no. 2, pp. 102–106, Jan. 2024, doi: 10.52625/j-agr.v19i2.297.
- [8] Supijatno, F. Adrian, dan A. Junaedi, "Analisis Pengelolaan Pemupukan Tanaman Kelapa Sawit (*Elaeis guineensis* Jacq.) Sebagai Bagian dari Best Management Practices Di Kebun Mesuji," *Buletin Agrohorti*, vol. 13, no. 1, hal. 40–47, 2025, doi: 10.29244/agrob.v13i1.60817.
- [9] E. Edric and S. P. Tamba, "Prediksi Penyakit Gagal Jantung dengan Menggunakan Random Forest," *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 176–181, Feb. 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445.
- [10] K. Aqbar and R. A. Supono, "Performance Analysis of LSTM and XGBoost Models Optimization in Forecasting Crude Palm Oil (CPO) Production at Palm Oil Mill (POM)," *Int. J. Comput. Appl.*, vol. 185, no. 17, pp. 37–44, Jun. 2023, doi: 10.5120/ijca2023922890.
- [11] K. Khaidir, F. Fadhliani, Z. Wirda, and A. Ramadhani, "Model Prediksi Produksi Pertanian Berbasis Machine Learning dan Data Lapangan," *SISFO: Jurnal Ilmiah Sistem Informasi*, vol. 9, no. 2, pp. 172–182, Oct. 2025, doi: 10.29103/v9i2.26015.
- [12] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development," *Clin. Transl. Sci.*, vol. 17, no. 11, art. no. e70056, Nov. 2024, doi: 10.1111/cts.70056.
- [13] A. Prasetyo, S. Salahuddin, and A. Amirullah, "Prediksi Produksi Kelapa Sawit Menggunakan Metode Regresi Linier Berganda," *Jurnal Infomedia*, vol. 6, no. 2, p. 76, Dec. 2021, doi: 10.30811/jim.v6i2.2343.
- [14] B. Wulan Sari and D. Prabowo, "Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, Dan XGBoost," *Intellect : Indonesian Journal of Learning and Technological Innovation*, vol. 4, no. 1, pp. 42–51, Jun. 2025, doi: 10.57255/intellect.v4i1.1385.
- [15] M. Nalenz, J. Rodemann, and T. Augustin, "Learning De-Biased Regression Trees and Forests from Complex Samples," *Mach. Learn.*, vol. 113, no. 6, pp. 3379–3398, Jun. 2024, doi: 10.1007/s10994-023-06439-1.
- [16] C. Lukita, L. D. Bakti, U. Rusilowati, A. Sutarman, and U. Rahardja, "Predictive and Analytics Using Data Mining and Machine Learning for Customer Churn Prediction," *Journal of Applied Data Sciences*, vol. 4, no. 4, pp. 454–465, Dec. 2023, doi: 10.47738/jads.v4i4.131.
- [17] R. I. Nthangeni, C. Sigauke, T. Ravele, and T. H. Tshisikhawe, "Enhancing Short-Term Wind Energy Forecasting with XGBoost and Conformal Prediction for Robust Uncertainty Quantification," *Computation*, vol. 14, no. 3, p. 56, Mar. 2026, doi: 10.3390/computation14030056.
- [18] A. Meiseles and L. Rokach, "Iterative Feature Exclusion (IFX): Mitigating Feature Starvation in Gradient Boosted Decision Trees," *Knowl. Based. Syst.*, vol. 289, art. no. 111546, Apr. 2024, doi: 10.1016/j.knosys.2024.111546.
- [19] M. Azad, T. H. Nehal, and M. Moshkov, "A Novel Ensemble Learning Method Using Majority Based Voting of Multiple Selective Decision Trees," *Computing*, vol. 107, no. 1, p. 42, Jan. 2025, doi: 10.1007/s00607-024-01394-8.
- [20] U. Ahmed et al., "Hybrid Bagging and Boosting with SHAP Based Feature Selection for Enhanced Predictive Modeling in Intrusion Detection Systems," *Sci. Rep.*, vol. 14, no. 1, p. 30532, Dec. 2024, doi: 10.1038/s41598-024-81151-1.