JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Perbandingan Random Forest, XGBoost, dan LightGBM untuk Prediksi Produksi Padi di Sumatera

Namariq Maspupah[1*], Deasy Purwaningtias[2], Ali Mustopa[3]

[1,2,3]Program Studi Informatika, Fakultas Teknik dan Informatika,Universitas Bina Sarana Informatika,Pontianak, Indonesia
Email: [1*]namariqmaspupah@gmail.com, [2]deasy.dwg@bsi.ac.id, [3]alimustopa.aop@bsi.ac.id
(*Email Corresponding Author: namariqmaspupah@gmail.com)
Received: 29 Juni 2026 | Revision: 4 Juli 2026 | Accepted: 8 Juli 2026

Abstrak

Produksi padi di Sumatera yang memiliki peran strategis dalam mendukung ketahanan pangan nasional, namun fluktuasi yang dipengaruhi oleh faktor iklim dan agronomis kadang menyulitkan proses perencanaan distribusi pangan secara akurat dan tepat. Dalam penelitian ini memiliki tujuan untuk membangun dan membandingkan model prediksi produksi padi di delapan daerah di provinsi Sumatera menggunakan tiga algoritma machine learning, yaitu Random Forest, XGBoost, dan LightGBM, yang masing-masing dioptimalkan menggunakan GridSearchCV dengan validasi silang 5-fold. Dataset yang digunakan bersumber dari Badan Pusat Statistik (BPS) dengan periode 1993–2020, mencakup 224 data dari variabel luas panen, curah hujan, kelembapan, dan suhu rata-rata. Interpretasi model dilakukan menggunakan metode SHAP (SHapley Additive exPlanations) untuk menganalisis kontribusi setiap fitur terhadap hasil prediksi. Hasil eksperimen menunjukkan bahwa XGBoost dengan parameter terbaik (learning_rate=0.05, max_depth=5, n_estimators=300, subsample=0.8) menghasilkan performa tertinggi dengan nilai MAE sebesar 105.258 ton, RMSE sebesar 151.883 ton, dan R^2 sebesar 0,9739. Random Forest menghasilkan R^2 sebesar 0,9580, sedangkan LightGBM menghasilkan R^2 sebesar 0,9519. Analisis SHAP pada ketiga model secara konsisten menunjukkan bahwa luas panen merupakan fitur paling dominan dalam memengaruhi prediksi produksi padi, diikuti oleh variabel tahun, kelembapan, curah hujan, suhu rata-rata, dan provinsi. Dalam penelitian ini memberikan kontribusi berupa kerangka perbandingan komprehensif antara tiga algoritma ensemble terbaik dengan integrasi Explainable AI pada konteks prediksi produksi padi di Sumatera.

Kata Kunci: Random Forest, XGBoost, LightGBM, Prediksi Produksi Padi, SHAP Value, Hyperparameter Tuning, Sumatera

Abstract

Rice production in Sumatra plays a strategic role in supporting national food security; however, fluctuations influenced by climatic and agronomic factors often complicate the process of accurate and timely food distribution planning. This study aims to develop and compare rice production prediction models across eight provinces in Sumatra using three machine learning algorithms, namely Random Forest, XGBoost, and LightGBM, each optimized using GridSearchCV with 5-fold cross-validation. The dataset was obtained from the Central Bureau of Statistics (BPS) covering the period 1993–2020, consisting of 224 records with variables including harvested area, rainfall, humidity, and average temperature. Model interpretation was performed using the SHAP (SHapley Additive exPlanations) method to analyze the contribution of each feature to the prediction outcomes. Experimental results show that XGBoost with optimal parameters (learning_rate=0.05, max_depth=5, n_estimators=300, subsample=0.8) achieved the highest performance with an MAE of 105,258 tons, RMSE of 151,883 tons, and R^2 of 0.9739. Random Forest yielded an R^2 of 0.9580, while LightGBM achieved 0.9519. SHAP analysis across all three models consistently identified harvested area as the most dominant feature influencing rice production predictions, followed by year, humidity, rainfall, average temperature, and province. This study contributes a comprehensive comparison framework among three top ensemble algorithms with integrated Explainable AI in the context of rice production prediction in Sumatra.

***Keywords**: Random Forest, XGBoost, LightGBM, Rice Production Prediction, SHAP Value, Hyperparameter Tuning, Sumatra*

1. PENDAHULUAN

Ketahanan pangan merupakan salah satu prioritas utama dalam pembangunan nasional Indonesia. Berdasarkan Undang-Undang Nomor 18 Tahun 2012 tentang Pangan, ketahanan pangan didefinisikan sebagai kondisi terpenuhinya kebutuhan pangan bagi seluruh warga negara yang tercermin dari ketersediaan pangan yang cukup, baik dari segi jumlah maupun mutu, aman, bergizi, merata, dan terjangkau. Padi (Oryza sativa L.) sebagai komoditas pangan utama Indonesia menjadi tumpuan bagi lebih dari 270 juta penduduk dalam memenuhi kebutuhan karbohidrat harian. Pulau Sumatera memberikan kontribusi signifikan terhadap total produksi padi nasional, dengan delapan provinsi utama yakni Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Lampung yang memiliki karakteristik agroklimat beragam dan kapasitas produksi yang berbeda-beda. Selain memengaruhi produksi pertanian, kondisi iklim juga berpengaruh terhadap pengelolaan sumber daya air, produktivitas tanaman, serta ketahanan pangan sehingga diperlukan model prediksi yang mampu menangkap hubungan yang kompleks antarvariabel[1].

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Namun demikian, produksi padi di Sumatera menghadapi tantangan berupa fluktuasi yang sulit diprediksi. Faktor iklim seperti perubahan pola curah hujan, peningkatan suhu rata-rata akibat pemanasan global, serta variabilitas kelembapan udara secara langsung memengaruhi produktivitas tanaman padi. Di sisi lain, faktor agronomis seperti luas lahan panen yang berfluktuasi akibat alih fungsi lahan turut berkontribusi terhadap ketidakpastian produksi. Ketidakmampuan dalam memprediksi produksi padi secara akurat berdampak pada perencanaan distribusi pangan yang tidak optimal, potensi krisis pangan lokal, serta ketidakefisienan kebijakan subsidi pertanian. Oleh karena itu, diperlukan sebuah sistem prediksi produksi padi yang andal berbasis data historis dan pendekatan komputasional yang mutakhir. Selain menghasilkan prediksi yang akurat, model machine learning juga perlu memiliki tingkat interpretabilitas yang baik agar faktor-faktor yang memengaruhi hasil prediksi dapat dipahami oleh pengguna. Interpretasi model menjadi penting terutama pada bidang pertanian karena dapat membantu peneliti maupun pengambil kebijakan dalam mengidentifikasi variabel yang paling berpengaruh terhadap produktivitas tanaman[2]. Penelitian sebelumnya menunjukkan bahwa pendekatan *Explainable Machine Learning* mampu meningkatkan transparansi model sekaligus memberikan informasi mengenai kontribusi setiap variabel terhadap hasil prediksi.[3]

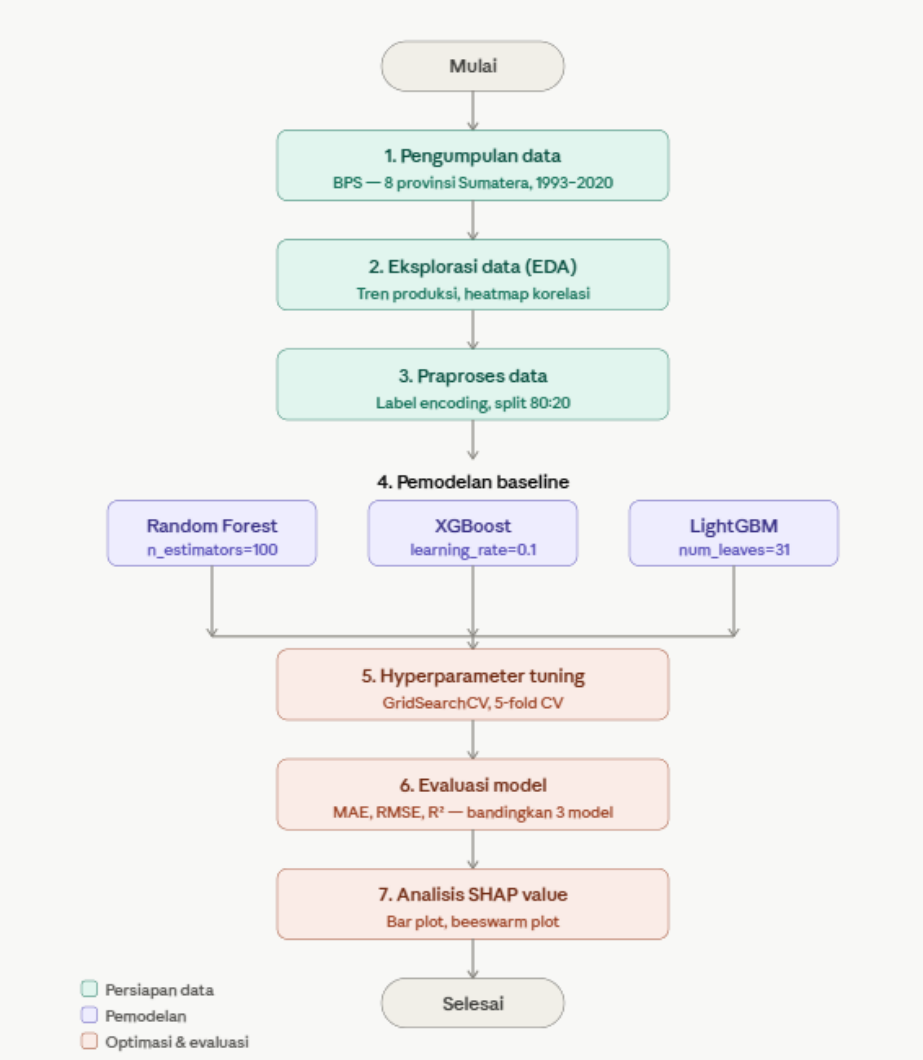
Perkembangan algoritma machine learning telah banyak dimanfaatkan untuk menyelesaikan permasalahan prediksi karena mampu memodelkan hubungan nonlinier yang sulit dijelaskan menggunakan pendekatan statistik konvensional. Beberapa penelitian sebelumnya telah mengeksplorasi penerapan machine learning dalam prediksi produksi pertanian. Adhany et al. [4]menerapkan algoritma Long Short-Term Memory (LSTM) untuk memprediksi produksi padi di Kota Lubuklinggau menggunakan data bulanan periode 2019–2024 dan memperoleh nilai MAPE sebesar 4,44% yang tergolong sangat baik, namun penelitian tersebut hanya mencakup satu kota di Sumatera Selatan dan tidak mengintegrasikan metode Explainable AI. Sylvia et al. [5] menggunakan Support Vector Regression (SVR) pada beberapa provinsi di Sumatera dengan variabel iklim sebagai prediktor, namun kinerja model terbatas akibat sensitivitas SVR terhadap skala data dan ketiadaan mekanisme interpretasi model. Satria et al. [6]membandingkan enam algoritma machine learning Random Forest, Decision Tree, Gradient Boosting, Extra Tree, SVM, dan ANN untuk memprediksi hasil panen tanaman pangan di Sumatera dan menemukan Extra Tree memperoleh R^2 tertinggi sebesar 0,968, namun penelitian tersebut tidak mengintegrasikan Hyperparameter Tuning maupun analisis SHAP. Al Anshori et al. [7] membangun model prediksi produksi padi di Sumatera menggunakan XGBoost berbasis framework CRISP-DM dengan dataset 224 observasi periode 1993–2020, namun penelitian tersebut hanya berfokus pada satu algoritma dan belum melakukan perbandingan lintas algoritma maupun integrasi Explainable AI. Khaidir et al. [2] membandingkan empat algoritma machine learning Random Forest, SVR, XGBoost, dan ANN untuk prediksi produksi pertanian menggunakan data lapangan di Aceh dengan XGBoost menghasilkan R^2 terbaik sebesar 0,89, namun penelitian tersebut tidak mencakup wilayah Sumatera secara menyeluruh dan tidak mengintegrasikan analisis SHAP.Meskipun berbagai penelitian telah berhasil meningkatkan akurasi prediksi menggunakan algoritma machine learning, sebagian besar penelitian masih berfokus pada evaluasi kinerja model tanpa memberikan penjelasan mengenai faktor-faktor yang memengaruhi hasil prediksi. Padahal, interpretabilitas model sangat penting untuk mendukung pengambilan keputusan pada sektor pertanian. Oleh karena itu penelitian ini menganalisis penelitian-penelitian terdahulu mengidentifikasi tiga celah penelitian (gap) yang signifikan. Pertama, belum ada penelitian yang secara komprehensif membandingkan tiga algoritma ensemble modern Random Forest, XGBoost, dan LightGBM secara bersamaan dalam konteks prediksi produksi padi di seluruh wilayah Sumatera. Kedua, integrasi antara Hyperparameter Tuning berbasis GridSearchCV dan Explainable AI berbasis SHAP pada ketiga model secara simultan belum pernah dilakukan. Ketiga, penelitian terdahulu umumnya menggunakan rentang data yang lebih pendek dan belum memanfaatkan keseluruhan variabel iklim serta agronomis secara bersama-sama. Berdasarkan identifikasi gap tersebut, penelitian ini bertujuan untuk: (1) membangun model prediksi produksi padi di Sumatera menggunakan tiga algoritma machine learning (Random Forest, XGBoost, dan LightGBM); (2) mengoptimalkan performa masing-masing model melalui Hyperparameter Tuning menggunakan GridSearchCV; (3) membandingkan performa ketiga model berdasarkan metrik evaluasi MAE, RMSE, dan R^2; serta (4) menginterpretasikan model terbaik menggunakan SHAP Value untuk mengidentifikasi fitur yang paling berpengaruh terhadap prediksi produksi padi. Hasil penelitian ini diharapkan dapat memberikan kontribusi ilmiah berupa kerangka metodologi prediksi produksi padi yang akurat dan dapat diinterpretasikan, serta menjadi acuan bagi pengambil kebijakan dalam perencanaan ketahanan pangan di wilayah Sumatera.

2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini melewati enam tahapan sistematis yang saling berkaitan. Tahapan yang pertama pengumpulan data, yaitu mengumpulkan data historis produksi padi dan variabel-variabel terkait dari sumber resmi Badan Pusat Statistik (BPS). Tahap kedua adalah eksplorasi data (Exploratory Data Analysis/EDA), yang mencakup pemeriksaan struktur data, identifikasi nilai kosong, analisis statistik deskriptif, dan visualisasi tren produksi per provinsi. Tahap ketiga adalah praproses data, meliputi encoding variabel kategorikal dan pembagian dataset. Tahap keempat adalah pembangunan model baseline dari tiga algoritma yang dipilih. Tahap kelima adalah optimasi model melalui Hyperparameter Tuning menggunakan GridSearchCV. Tahap keenam adalah evaluasi dan interpretasi model menggunakan metrik kuantitatif dan analisis SHAP Value.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378



Gambar 1. Alur Penelitian

2.2 Algoritma Machine Learning

Penelitian ini mengimplementasikan tiga algoritma ensemble berbasis pohon keputusan. Random Forest adalah metode ensemble yang membangun sejumlah pohon keputusan secara paralel menggunakan teknik bootstrap aggregating (bagging), kemudian menggabungkan hasil prediksi seluruh pohon melalui rata-rata untuk tugas regresi[8]. XGBoost (Extreme Gradient Boosting) adalah implementasi gradient boosting yang membangun pohon secara sekuensial, di mana setiap pohon baru dilatih untuk memperbaiki residual pohon sebelumnya, dengan dukungan regularisasi L1 dan L2 yang terintegrasi untuk mencegah overfit[9]. LightGBM (Light Gradient Boosting Machine) adalah framework gradient boosting yang dikembangkan oleh Microsoft dengan fokus pada efisiensi komputasi melalui teknik *Gradient-based One-Side* Sampling (GOSS) dan Exclusive Feature Bundling (EFB), serta menggunakan strategi pertumbuhan pohon *leaf-wise* yang menghasilkan akurasi lebih tinggi dengan iterasi lebih sedikit[10].

2.3 Hyperparameter Tuning

Hyperparameter tuning dilakukan menggunakan metode GridSearchCV dengan validasi silang lima lipatan (*5-fold cross-validation*) yang diimplementasikan pada pustaka Scikit-learn[11]. Pada algoritma Random Forest**,** ruang pencarian (*search space*) meliputi parameter *n_estimators* = {50, 100, 200}, *max_depth* = {None, 5, 10, 20}, dan *min_samples_split* = {2, 5, 10}. Hasil pencarian menunjukkan bahwa kombinasi parameter terbaik adalah *n_estimators* = 50, *max_depth* = None, dan *min_samples_split* = 2. Selanjutnya, pada algoritma XGBoost**,** parameter yang dioptimalkan meliputi *n_estimators* = {100, 200, 300}, *max_depth* = {3, 5, 7}, *learning_rate* = {0.05, 0.1, 0.2}, dan *subsample* = {0.8, 1.0}. Berdasarkan proses optimasi, diperoleh parameter terbaik yaitu *n_estimators* = 300, *max_depth* = 5, *learning_rate* = 0.05, dan *subsample* = 0.8. Sementara itu, pada algoritma LightGBM, parameter yang diuji meliputi *n_estimators* = {100, 200, 300}, *max_depth* = {3, 5, 7}, *learning_rate* = {0.05, 0.1, 0.2}, dan *num_leaves* = {31, 63, 127}. Hasil optimasi menunjukkan bahwa kombinasi parameter terbaik adalah *n_estimators* = 100, *max_depth* = 5, *learning_rate* = 0.05, dan *num_leaves* = 31.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

2.4 Metrik Evaluasi

Performa model dievaluasi menggunakan tiga metrik kuantitatif. Mean Absolute Error (MAE) mengukur rata-rata kesalahan prediksi secara absolut:

$$MAE = (1/n) \times \Sigma|yi - \hat{y}i| \quad (1)$$

Root Mean Square Error (RMSE) memberikan penalti lebih besar pada kesalahan besar dan lebih sensitif terhadap outlier:

$$RMSE = \sqrt{[(1/n) \times \Sigma(yi - \hat{y}i)^2]} \quad (2)$$

Koefisien Determinasi (R^2) mengukur proporsi variansi data yang dapat dijelaskan oleh model, dengan rentang 0 hingga 1 di mana nilai mendekati 1 menunjukkan model yang sangat baik:

$$R^2 = 1 - [\Sigma(yi - \hat{y}i)^2] / [\Sigma(yi - \bar{y})^2] \quad (3)$$

2.5 SHAP Value

Interpretasi model dilakukan menggunakan SHAP (SHapley Additive exPlanations) yang dikembangkan oleh Lundberg dan Lee berdasarkan teori permainan kooperatif[12]. SHAP menghitung kontribusi marginal setiap fitur terhadap prediksi model dengan nilai Shapley yang merupakan rata-rata tertimbang dari kontribusi fitur pada semua kemungkinan koalisi fitur. Nilai SHAP positif berarti fitur mendorong prediksi di atas nilai rata-rata (baseline), sedangkan nilai SHAP negatif berarti fitur menurunkan prediksi. Visualisasi dilakukan menggunakan SHAP Bar Plot untuk melihat kepentingan fitur secara global dan SHAP Beeswarm Plot untuk melihat arah dan distribusi pengaruh setiap fitur. Pendekatan Explainable Machine Learning memungkinkan model machine learning tidak hanya menghasilkan prediksi yang akurat, tetapi juga memberikan penjelasan mengenai kontribusi setiap variabel terhadap hasil prediksi sehingga proses pengambilan keputusan menjadi lebih transparan.[3]

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari dataset "Dataset Tanaman Padi Sumatera, Indonesia" yang tersedia pada platform Kaggle[13]. Dataset tersebut merupakan kompilasi data produksi padi dan luas panen yang bersumber dari Badan Pusat Statistik (BPS)**,** serta data curah hujan, kelembapan, dan suhu rata-rata yang berasal dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG)**,** mencakup delapan provinsi di Pulau Sumatera selama periode 1993–2020. Dataset memiliki total 224 baris data dengan 7 kolom variabel, tanpa adanya nilai kosong (missing values). Variabel yang digunakan terdiri atas variabel independen (fitur) yaitu Provinsi, Tahun, Luas Panen (hektar), Curah Hujan (mm), Kelembapan (%), dan Suhu Rata-rata (°C), serta variabel dependen (target) yaitu Produksi Padi (ton).

Statistik deskriptif menunjukkan bahwa rata-rata produksi padi adalah 1.679.701 ton dengan nilai minimum 429.300 ton dan maksimum 4.881.089 ton. Rata-rata luas panen adalah 374.350 hektar, rata-rata curah hujan adalah 2.452 mm/tahun, rata-rata kelembapan adalah 80,95%, dan rata-rata suhu adalah 26,80°C. Analisis korelasi menunjukkan bahwa luas panen memiliki korelasi tertinggi dengan produksi padi sebesar 0,91, sedangkan variabel iklim (curah hujan, kelembapan, suhu) memiliki korelasi yang relatif rendah terhadap produksi.

Tabel 1. Data Pertanian

Provinsi	Luas Panen (ha) 1993	...	Luas Panen (ha) 2020	Produksi (ton) 1993	...	Produksi (ton) 2020
Aceh	323.589	...	317.869	1.329.536	...	1.861.567
Sumatera Utara	754.569	...	388.591	2.918.152	...	2.076.280
Sumatera Barat	394.412	...	295.664	1.806.424	...	1.450.840
Riau	146.133	...	64.733	436.297	...	269.344
Jambi	199.431	...	84.773	607.529	...	374.376
Sumatera Selatan	439.895	...	551.321	1.409.559	...	2.696.877
Bengkulu	109.807	...	64.137	356.709	...	296.925
Lampung	433.078	...	545.149	1.646.900	...	2.604.913

Tabel 2. Data Iklim

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Provinsi	Curah Hujan (mm) 1993	...	Curah Hujan (mm) 2020	Kelembapan (%) 1993	...	Kelembapan (%) 2020	Suhu (°C) 1993	...	Suhu (°C) 2020
Aceh	1.627,0	...	1.619,2	82,00	...	80,82	26,06	...	25,41
Sumatera Utara	2.549,4	...	1.648,3	83,80	...	83,76	26,45	...	25,79
Sumatera Barat	5.110,0	...	4.072,7	89,77	...	85,22	24,70	...	26,26
Riau	2.738,4	...	2.584,9	86,37	...	81,80	26,00	...	25,93
Jambi	1.954,0	...	2.303,8	83,81	...	78,24	26,29	...	25,85
Sumatera Selatan	2.815,0	...	2.300,2	82,78	...	81,10	26,38	...	26,14
Bengkulu	4.150,0	...	4.144,0	85,50	...	80,18	25,95	...	25,84
Lampung	2.306,7	...	2.211,3	84,82	...	75,80	26,41	...	24,58

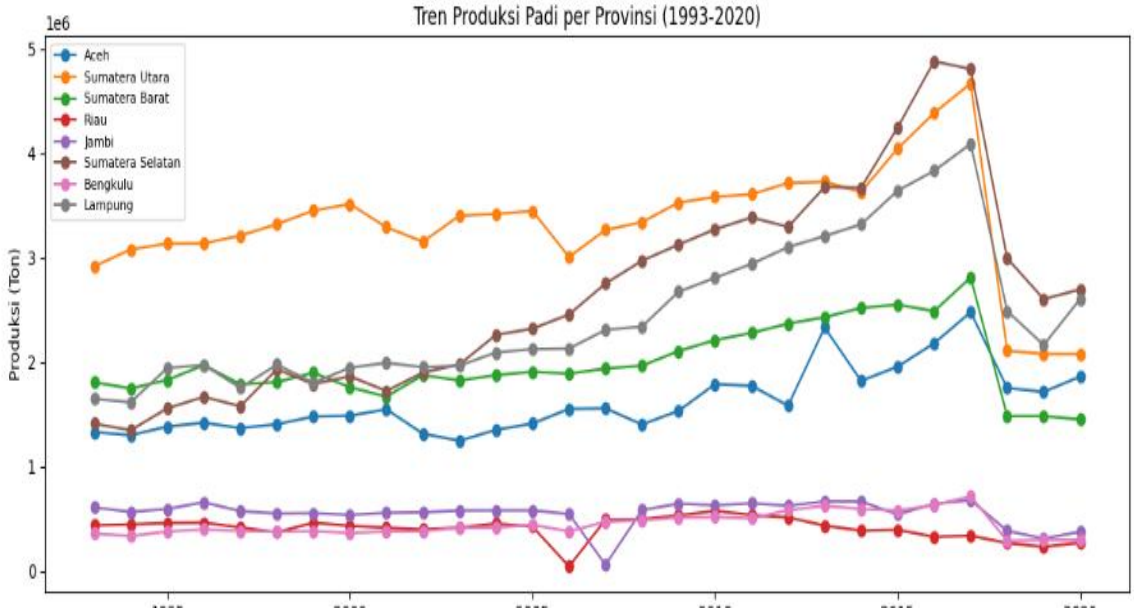
3.2 Praproses Data

Tahap praproses data mencakup dua langkah utama. Pertama, encoding variabel kategorikal dilakukan terhadap kolom Provinsi menggunakan Label Encoder, yang mengubah nama-nama provinsi menjadi representasi numerik agar dapat diproses oleh algoritma machine learning. Kedua, pembagian dataset dilakukan dengan proporsi 80% data pelatihan (179 data) dan 20% data pengujian (45 data) menggunakan fungsi train test split dengan parameter random state 42 untuk memastikan reprodusibilitas hasil.

3.3 Eksplorasi Data

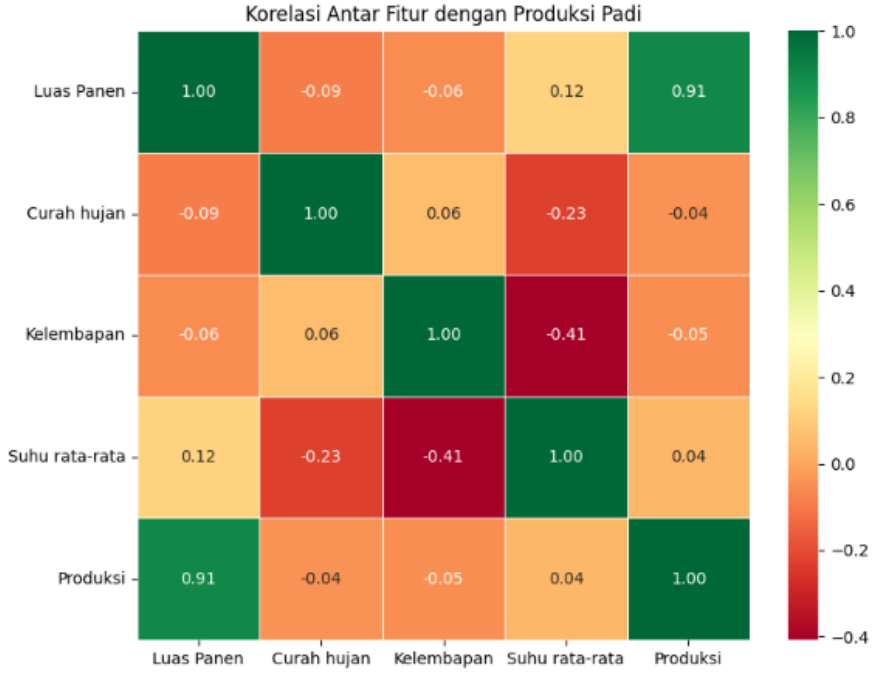
Analisis eksplorasi data menunjukkan bahwa dataset memiliki kualitas yang baik, ditandai dengan tidak adanya nilai kosong pada seluruh 224 entri data. Tren produksi padi per provinsi selama 1993–2020 memperlihatkan pola peningkatan secara umum, dengan Sumatera Utara sebagai provinsi dengan produksi tertinggi (rata-rata di atas 3 juta ton) diikuti oleh Sumatera Selatan dan Lampung, sementara Bengkulu dan Riau memiliki produksi terendah. Terdapat penurunan produksi yang signifikan pada tahun 2015–2016 di sebagian besar provinsi, yang kemungkinan berkaitan dengan fenomena El Niño yang menyebabkan kekeringan ekstrem.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378



Gambar 2. Tren Produksi Padi Per Provinsi (1993-2020)

Analisis korelasi antar fitur dengan variabel target (Produksi) mengungkap temuan yang penting. Luas Panen memiliki korelasi sangat tinggi dengan Produksi (r = 0,91), mengindikasikan bahwa luas area yang dipanen merupakan penentu utama volume produksi. Sebaliknya, variabel iklim menunjukkan korelasi rendah: Curah Hujan (r = -0,04), Kelembapan (r = -0,05), dan Suhu Rata-rata (r = 0,04).



Gambar 3. Korelasi Antar Fitur Dengan Produksi Padi

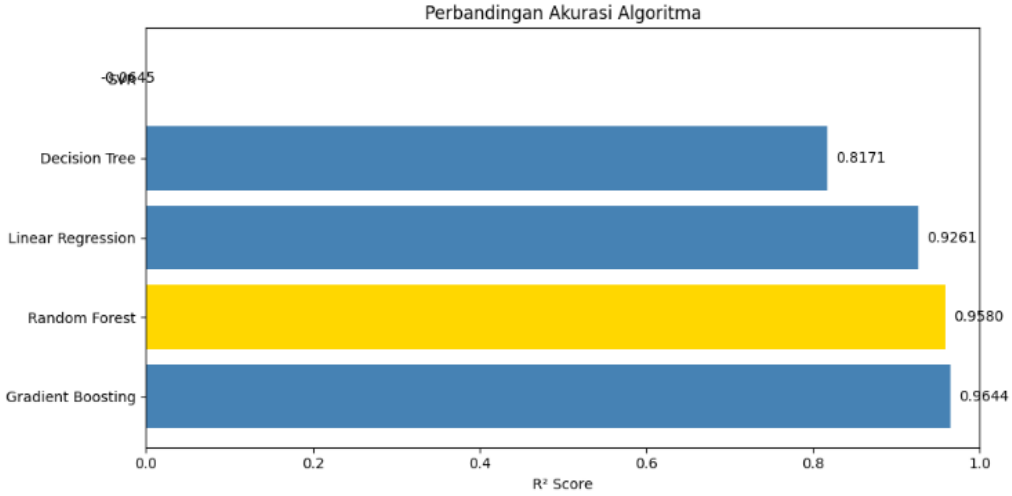
Hal ini tidak berarti variabel iklim tidak berpengaruh, melainkan pengaruhnya bersifat non-linear dan berinteraksi satu sama lain, yang justru dapat ditangkap dengan baik oleh algoritma ensemble berbasis pohon.

3.4 Perbandingan Model Baseline

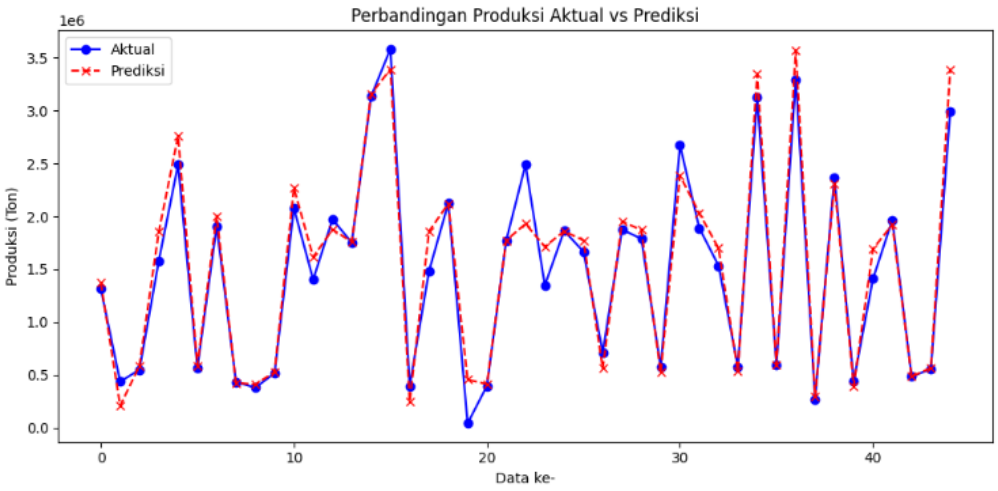
Sebelum dilakukan Hyperparameter Tuning, kelima algoritma yang diuji dalam tahap eksplorasi awal menghasilkan performa sebagai berikut. Gradient Boosting menempati posisi teratas dengan R^2 = 0,9644 dan MAE = 130.000 ton, diikuti Random Forest (R^2 = 0,9580, MAE = 135.772 ton), Linear Regression (R^2 = 0,9261, MAE = 191.004 ton),

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Decision Tree (R^2 = 0,8171, MAE = 195.369 ton), dan SVR yang menghasilkan R^2 negatif (-0,0645), menandakan ketidakcocokan algoritma tersebut tanpa proses normalisasi data terlebih dahulu.



Gambar 4. Perbandingan Akurasi Algoritma



Gambar 5. Perbandingan Produksi Aktual Vs Produksi

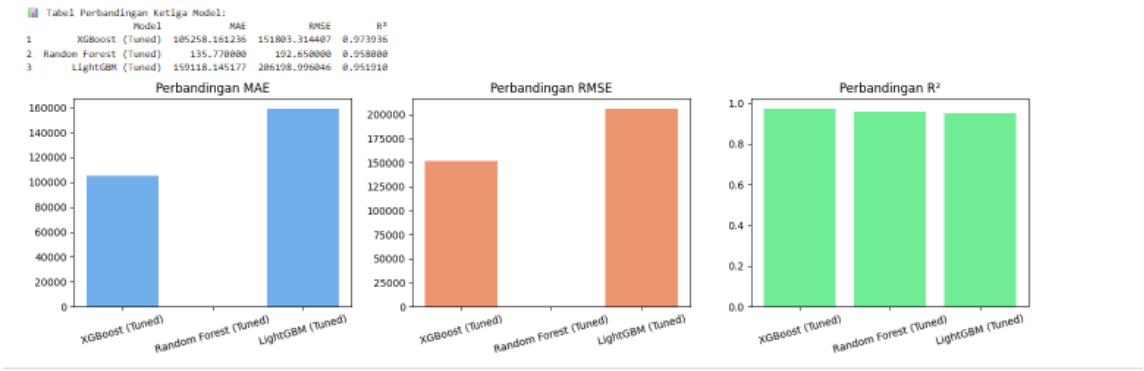
3.5 Perbandingan Model Setelah Parameter Tuning

Setelah optimasi dengan GridSearchCV, perbandingan performa ketiga model utama disajikan pada Tabel 3.

Tabel 3. Perbandingan Performa Model Setelah Hyperparameter Tuning

No	Model	MAE (ton)	RMSE (ton)	R^2
1	XGBoost (Tuned)	105.258	151.883	0,9739
2	Random Forest (Tuned)	135.770	192.650	0,9580
3	LightGBM (Tuned)	159.118	206.199	0,9519

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378



***Gambar 6**. Perbandingan 3 Model MAE,RMSE dan R^2*

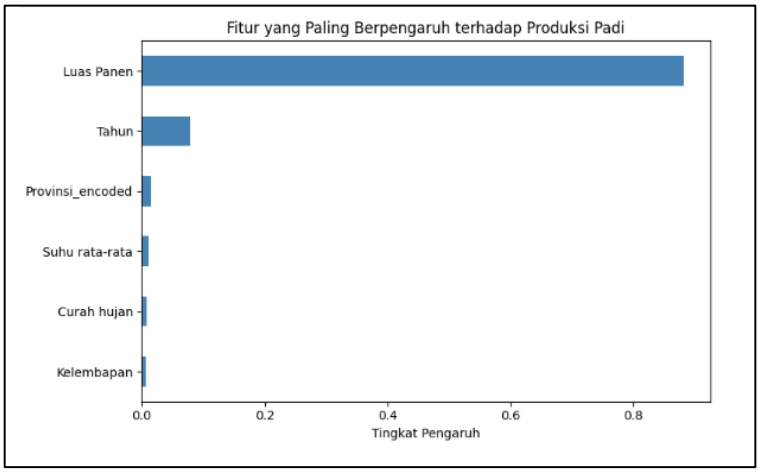
Hasil pada Tabel 1 menunjukkan bahwa XGBoost dengan Hyperparameter Tuning menghasilkan performa terbaik di antara ketiga model, dengan nilai R^2 sebesar 0,9739 yang berarti model mampu menjelaskan 97,39% variansi data produksi padi. Nilai MAE sebesar 105.258 ton menunjukkan rata-rata kesalahan prediksi sekitar 105 ribu ton, yang merupakan tingkat akurasi yang sangat baik mengingat rata-rata produksi adalah 1,68 juta ton (kesalahan relatif sekitar 6,27%).

XGBoost unggul dibandingkan Random Forest dengan selisih R^2 sebesar 0,0159 poin dan penurunan MAE sebesar 30.512 ton (22,45%). Keunggulan XGBoost ini disebabkan oleh mekanisme boosting sekuensial yang secara iteratif memperbaiki kesalahan prediksi sebelumnya, didukung regularisasi L1 dan L2 yang mencegah overfitting secara efektif. LightGBM menghasilkan performa sedikit di bawah Random Forest, dengan R^2 = 0,9519. Hal ini kemungkinan disebabkan oleh ukuran dataset yang relatif kecil (224 data), di mana keunggulan efisiensi LightGBM untuk data besar belum dapat dimanfaatkan secara optimal, dan strategi pertumbuhan pohon leaf-wise berpotensi mengalami overfitting pada dataset kecil jika tidak dikontrol dengan ketat.

Perbandingan antara performa sebelum dan sesudah hyperparameter tuning pada Random Forest menunjukkan perubahan yang sangat kecil. Nilai MAE berubah dari 135.772 ton menjadi 135.770 ton**,** sedangkan RMSE berubah dari 192.647 ton menjadi 192.650 ton**,** dengan nilai R^2 tetap berada pada kisaran 0,9580**.** Hasil ini menunjukkan bahwa parameter bawaan (default) Random Forest telah memberikan performa yang hampir optimal pada dataset yang digunakan. Temuan ini sejalan dengan Probst et al. [14]yang menjelaskan bahwa Random Forest umumnya memiliki performa yang baik dengan parameter awal, sedangkan peningkatan melalui hyperparameter tuning sangat bergantung pada karakteristik dataset.

3.6 Analisis SHAP Value

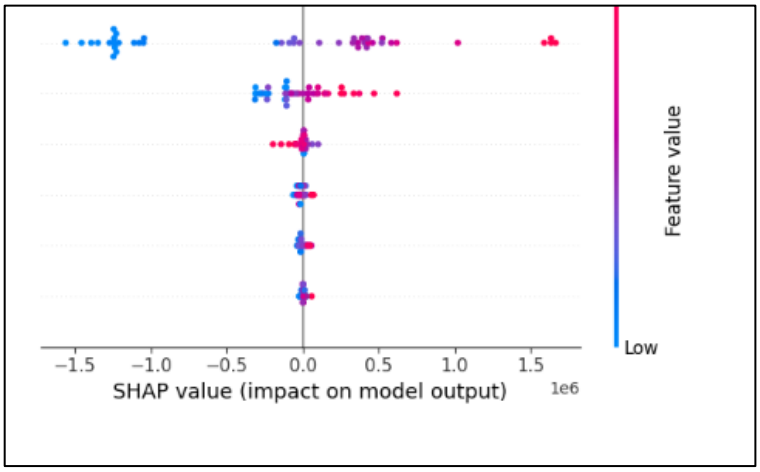
Analisis menggunakan SHAP dilakukan untuk mengevaluasi kontribusi masing-masing variabel terhadap hasil prediksi secara global maupun lokal sehingga hubungan antara variabel input dan hasil prediksi dapat dipahami dengan lebih baik. Analisis SHAP juga dilakukan pada ketiga model untuk mengidentifikasi fitur-fitur yang paling berpengaruh terhadap prediksi produksi padi. Secara konsisten di ketiga model, Luas Panen merupakan fitur dengan nilai SHAP mean absolut tertinggi, jauh melampaui fitur lainnya. Hal ini sejalan dengan temuan analisis korelasi (r = 0,91) dan mengkonfirmasi bahwa volume produksi padi sangat ditentukan oleh seberapa besar area yang berhasil dipanen.



Gambar 7. SHAP Bar Plot – Random Forest

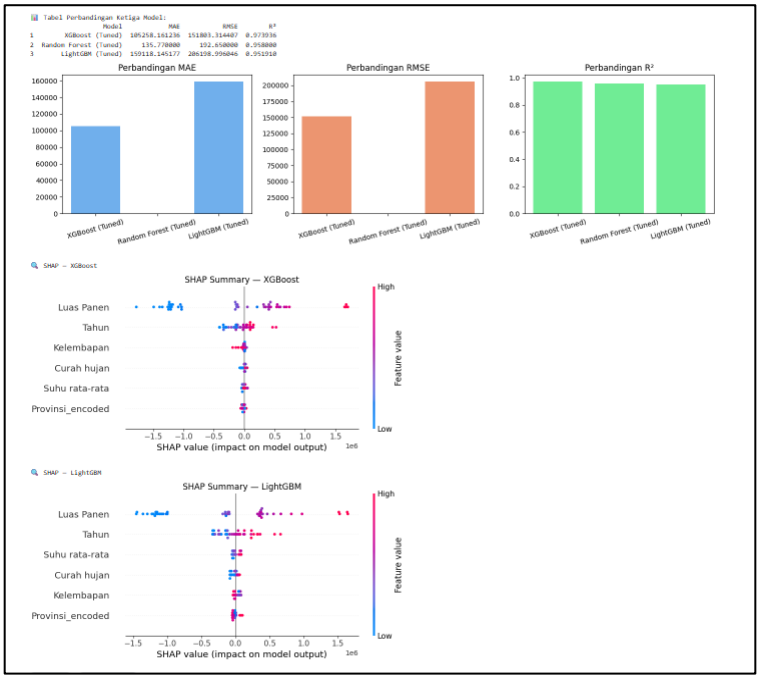
JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Pada model XGBoost, urutan kepentingan fitur berdasarkan SHAP adalah: (1) Luas Panen, (2) Tahun, (3) Kelembapan, (4) Curah Hujan, (5) Suhu rata-rata, (6) Provinsi_encoded. SHAP Beeswarm Plot pada XGBoost menunjukkan bahwa nilai Luas Panen yang tinggi (titik berwarna merah) memiliki nilai SHAP positif yang besar, yang berarti luas panen yang luas mendorong prediksi produksi ke nilai lebih tinggi, dan sebaliknya. Hasil ini menunjukkan bahwa metode Explainable Machine Learning mampu memberikan interpretasi terhadap model prediksi sehingga faktor-faktor yang paling berpengaruh dapat diidentifikasi dan dimanfaatkan sebagai dasar dalam penyusunan kebijakan di sektor pertanian. Fitur Tahun menunjukkan pola SHAP positif untuk nilai tahun yang lebih baru, mencerminkan tren peningkatan produktivitas dari waktu ke waktu akibat kemajuan teknologi pertanian.



Gambar 8. SHAP Beeswarm Plot – Random Forest

Pada model LightGBM, urutan kepentingan fitur sedikit berbeda: (1) Luas Panen, (2) Tahun, (3) Suhu rata-rata, (4) Curah Hujan, (5) Kelembapan, (6) Provinsi_encoded. Posisi Suhu rata-rata yang lebih tinggi pada LightGBM dibandingkan XGBoost mengindikasikan bahwa LightGBM lebih sensitif dalam menangkap pengaruh suhu terhadap produktivitas padi. Pada model Random Forest, urutan fitur adalah: (1) Luas Panen, (2) Tahun, (3) Kelembapan, (4) Provinsi_encoded, (5) Suhu rata-rata, (6) Curah Hujan.



Gambar 9. SHAP Summary - XGBoost dan LightGBM

Temuan konsisten dari analisis SHAP lintas ketiga model mengimplikasikan bahwa kebijakan peningkatan produksi padi di Sumatera harus memprioritaskan perluasan dan pemertahanan lahan sawah produktif sebagai intervensi utama. Variabel iklim meskipun memiliki nilai SHAP lebih rendah, tetap berperan dalam menentukan variabilitas produksi antar-tahun, sehingga sistem pemantauan iklim berbasis teknologi perlu terus dikembangkan. Selain menghasilkan model dengan performa prediksi yang baik, penelitian ini juga menunjukkan bahwa penerapan *Explainable*

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

Machine Learning melalui metode SHAP dapat meningkatkan transparansi model dengan menjelaskan kontribusi setiap variabel terhadap hasil prediksi.

3.7 Prediksi Produksi Padi 2021-2025

Model XGBoost terbaik digunakan untuk memprediksi produksi padi selama periode 2021–2025 berdasarkan rata-rata historis fitur per provinsi. Hasil prediksi menunjukkan Sumatera Utara diproyeksikan mempertahankan produksi tertinggi di kisaran 4,16 juta ton per tahun, diikuti Sumatera Selatan (4,16 juta ton), Sumatera Barat (2,61 juta ton), Lampung (2,59 juta ton), Aceh (1,77 juta ton), Jambi (0,61 juta ton), Riau (0,45 juta ton), dan Bengkulu (0,44 juta ton). Prediksi yang konstan antar tahun pada periode proyeksi mencerminkan penggunaan nilai rata-rata historis sebagai input fitur, sehingga model merepresentasikan kondisi produksi rata-rata jika kondisi pertanian tetap seperti kondisi historis.

4. KESIMPULAN

Penelitian ini telah berhasil membangun dan membandingkan tiga model machine learning Random Forest, XGBoost, dan LightGBM untuk prediksi produksi padi di delapan provinsi Sumatera menggunakan data historis BPS periode 1993–2020. Seluruh model dioptimalkan menggunakan GridSearchCV dengan validasi silang 5-fold dan diinterpretasikan menggunakan metode SHAP Value sebagai pendekatan Explainable AI. Berdasarkan hasil eksperimen, diperoleh kesimpulan bahwa XGBoost merupakan algoritma terbaik untuk kasus prediksi produksi padi di Sumatera, dengan nilai R^2 sebesar 0,9739, MAE sebesar 105.258 ton, dan RMSE sebesar 151.883 ton setelah *Hyperparameter Tuning* dengan *parameter optimal learning*_rate 0.05, max depth 5, nestimators 300, dan subsample 0.8. Hasil ini mengungguli Random Forest (R^2 = 0,9580) dan LightGBM (R^2 = 0,9519), menunjukkan bahwa mekanisme boosting sekuensial dengan regularisasi yang kuat pada XGBoost lebih efektif dalam menangkap pola produksi padi di wilayah Sumatera dibandingkan pendekatan bagging maupun *leaf-wise boosting*. Analisis SHAP secara konsisten mengidentifikasi Luas Panen sebagai fitur paling dominan di ketiga model, dengan nilai SHAP mean absolut yang jauh melampaui fitur lainnya, diikuti oleh Tahun sebagai representasi tren temporal peningkatan produktivitas. Variabel iklim (Kelembapan, Curah Hujan, dan Suhu Rata-rata) memberikan kontribusi yang lebih kecil namun tetap relevan dalam menjelaskan variabilitas produksi antar-tahun. Temuan ini mengimplikasikan bahwa kebijakan peningkatan produksi padi di Sumatera harus memprioritaskan perlindungan dan perluasan lahan sawah produktif, sekaligus mengembangkan sistem pemantauan iklim untuk mengantisipasi dampak perubahan iklim terhadap produktivitas pertanian. Untuk penelitian selanjutnya, disarankan untuk menambahkan variabel fitur yang lebih lengkap seperti indeks vegetasi (NDVI), penggunaan pupuk, dan ketersediaan air irigasi; memperluas cakupan data hingga periode yang lebih terkini; mengeksplorasi teknik tuning yang lebih canggih seperti *Bayesian Optimization;* serta mengembangkan model prediksi berbasis data spasial untuk analisis yang lebih granular pada tingkat kabupaten atau kecamatan.

REFERENCES

[1] I. Hapsari and S. Pandya Wisesa, “Evaluasi Model Prediksi Curah Hujan Berbasis Machine Learning di Kota Bandung,” *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 11, no. 2, pp. 136–143, Sep. 2025, doi: 10.25077/teknosi.v11i2.2025.136-143.

[2] 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.

[3] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” Jan. 01, 2021, *MDPI AG*. doi: 10.3390/e23010018.

[4] P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, and S. Zeger, “Springer Series in Statistics.” [Online]. Available: http://www.springer.com/series/692

[5] P. C. Adhany, C. Wulandari, B. Intan, and B. Santoso, “Prediksi Padi Menggunakan Algoritma Long Short Term Memory,” *Journal of Informatics Management and Information Technology*, vol. 5, no. 2, pp. 120–127, 2025, doi: 10.47065/jimat.v5i2.496.

[6] H. Purnomo, R. Estian Pambudi, O. Arifin, F. Kurniawan Ikhsan, P. Negeri Lampung, and I. Darmajaya Bandar Lampung, “Analisis Prediksi Produksi Tanaman Padi Berdasarkan Variabel Iklim Menggunakan Support Vector Regression.” [Online]. Available: https://jti.aisyahuniversity.ac.id/index.php/AJIEE

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1541-1551
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1378

[7] A. Satria, R. Maulida Badri, I. Safitri, and H. Artikel, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digital Transformation Technology*, vol. 3, no. 2, pp. 389–398, Sep. 2023, doi: 10.47709/DIGITECH.V3I2.2852.

[8] M. Ridwan Al Anshori *et al.*, "PREDIKSI PRODUKSI PADI DI PULAU SUMATERA MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST) BERBASIS FRAMEWORK CRISP-DM: STUDI DATA HISTORIS 1993 - 2020," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 10, no. 3, pp. 4567–4574, May 2026, doi: 10.36040/JATI.V10I3.18300.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785;CSUBTYPE:STRING:CONFERENCE.

[11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, Long Beach, California, USA: Curran Associates, Inc., 2017, pp. 3146–3154.

[12] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, I. ; L. U. V. ; B. S. ; W. H. ; F. R. ; V. S. ; G. R. Guyon, Ed., Long Beach, California, USA: Curran Associates, Inc., 2017, pp. 5765–4774.

[14] Dataset Tanaman Padi Sumatera, Indonesia. Available: https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera Accessed: July 4, 2026.

[15] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1301, May 2019, doi: 10.1002/WIDM.1301.