PREDIKSI PRODUKSI PADI DI PULAU SUMATERA MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST) BERBASIS FRAMEWORK CRISP-DM: STUDI DATA HISTORIS 1993 - 2020

Mochamad Ridwan Al Anshori, Apriade Voutama, Yuyun Umaidah
Sistem Informasi, Universitas Singaperbangsa Karawang
Jl. H.S. Ronggowaluyo, Telukjambe Timur, Kabupaten Karawang, Jawa Barat, Indonesia
ridwan02112002@gmail.com

ABSTRAK

Pulau Sumatera memegang peran strategis sebagai salah satu lumbung pangan nasional Indonesia dengan kontribusi produksi padi yang signifikan. Fluktuasi produksi yang tidak linear akibat variabilitas iklim dan keterbatasan data historis berskala regional menyulitkan perencanaan pangan yang akurat di tingkat provinsi. Penelitian ini bertujuan membangun model prediksi produksi padi untuk delapan provinsi di Pulau Sumatera menggunakan algoritma Extreme Gradient Boosting (XGBoost). Model dikembangkan mengikuti kerangka CRISP-DM dengan dataset 224 observasi periode 1993–2020. Variabel independen mencakup luas panen, curah hujan, kelembapan, dan suhu rata-rata. Konfigurasi model menerapkan regularisasi L1 (reg_alpha = 0,1) dan L2 (reg_lambda = 1,0) dengan max_depth = 3 untuk mengendalikan overfitting, serta dievaluasi melalui 5-Fold Cross Validation dan test set independen. Model XGBoost menghasilkan R^2 = 0,8734, MAE = 225.478 ton, dan RMSE = 334.613 ton pada test set, dengan R^2 cross-validation rata-rata 0,8434 ± 0,0267. Variabel provinsi menjadi prediktor dominan (feature importance = 54,56%), diikuti luas panen (33,25%). Generalization gap sebesar 0,1246 mengindikasikan overfitting parsial akibat keterbatasan jumlah data.

Kata kunci : *XGBoost, prediksi produksi padi, Pulau Sumatera, CRISP-DM, K-Fold Cross Validation, overfitting*

1. PENDAHULUAN

Padi (Oryza sativa L.) merupakan komoditas pangan strategis yang menjadi sumber kalori utama bagi lebih dari 70% penduduk Indonesia. Pulau Sumatera, sebagai salah satu lumbung pangan terbesar di Indonesia, menyumbang kontribusi signifikan terhadap total produksi padi nasional melalui delapan provinsi dengan karakteristik agroekologis yang heterogen. Namun, fluktuasi produksi yang tidak linear dan tren penurunan lahan sawah akibat alih fungsi lahan menjadi persoalan mendasar yang perlu diantisipasi secara sistemik [1].

Variabilitas iklim yang semakin tidak terprediksi, khususnya fenomena *El Niño* dan *La Niña*, terbukti memberikan dampak signifikan terhadap fluktuasi produksi padi di Indonesia. Periode *El Niño* 2023–2024 secara nyata menyebabkan kekeringan di sentra produksi padi, menurunkan luas panen nasional, dan memaksa pemerintah mengimpor jutaan ton beras untuk menutup defisit [2]. Kondisi ini menegaskan permasalahan mendasar yang dihadapi sektor pertanian Sumatera: ketiadaan sistem prediksi produksi padi yang mampu mengantisipasi fluktuasi secara dini berbasis data iklim historis dan luas lahan tanam, khususnya yang mempertimbangkan heterogenitas antarprovinsi dan kekhususan iklim tropis kawasan.

Metode regresi linear konvensional terbukti tidak mampu menangkap hubungan non-linear antara variabel iklim dan hasil produksi padi, terutama saat terjadi anomali iklim [3]. Pendekatan *machine learning* yang lebih adaptif menjadi keharusan dalam pemodelan pertanian modern. Di antara berbagai algoritma *ensemble* yang tersedia, *Extreme Gradient Boosting* (XGBoost) unggul karena menggabungkan mekanisme regularisasi bawaan (L1 dan L2) yang secara eksplisit mengontrol kompleksitas model, menjadikannya cocok untuk dataset pertanian berukuran terbatas [4]. Keunggulan *XGBoost* juga telah dibuktikan dalam studi prediksi produksi padi tahunan di Bangladesh, di mana algoritma ini mengungguli model *ARIMA* secara konsisten [5], membuka peluang penerapan serupa pada konteks regional Asia Tenggara, termasuk Sumatera.

Berdasarkan permasalahan dan kesenjangan penelitian tersebut, penelitian ini bertujuan membangun model prediksi produksi padi berbasis *XGBoost* untuk delapan provinsi di Pulau Sumatera yang valid secara statistik dan dapat diimplementasikan sebagai alat bantu pengambilan keputusan dalam perencanaan produksi padi di tingkat regional. Untuk mencapai tujuan tersebut, seluruh proses pengembangan model mengikuti kerangka kerja *Cross-Industry Standard Process* for *Data Mining* (CRISP-DM) [6]. dengan variabel independen berupa luas panen, curah hujan, kelembapan, dan suhu rata-rata, menggunakan dataset historis delapan provinsi Sumatera periode 1993–2020.

2. TINJAUAN PUSTAKA

2.1. Dinamika Produksi Padi dan Variabilitas Iklim

Perubahan iklim memberikan tekanan ganda pada sektor pertanian padi di Indonesia: meningkatkan frekuensi dan intensitas kejadian cuaca ekstrem sekaligus menggeser kalender musim tanam yang menjadi acuan petani. Hal ini berdampak langsung pada fluktuasi produktivitas lahan sawah, terutama

pada provinsi yang masih bergantung pada curah hujan sebagai sumber irigasi utama.

[7] dalam proyeksi produksi padi Pulau Sumatera hingga 2045 menggunakan model iklim regional *CORDEX-SEA* menemukan bahwa di bawah skenario emisi tinggi, provinsi-provinsi di Sumatera bagian selatan akan mengalami penurunan produktivitas padi yang lebih signifikan dibandingkan wilayah utara akibat pergeseran pola curah hujan. Temuan ini menggarisbawahi pentingnya model prediksi yang mampu menangkap heterogenitas respons iklim antarprovinsi di Pulau Sumatera.

2.2. Algoritma Extreme Gradient Boosting (XGBoost)

XGBoost memiliki keistimewaan fundamental pada fungsi objektifnya yang secara eksplisit mengintegrasikan komponen regularisasi L1 dan L2, sehingga mampu mengontrol kompleksitas model secara langsung selama proses pelatihan:

$$L(\theta) = \Sigma_i \, l(\hat{y}_i, y_i) + \Sigma_k \, \Omega(f_k), \;\; \Omega(\mathrm{f}) = \gamma T + {}^{1}\!/_{2} \, \lambda ||w||^2 + \alpha ||w||_1 \quad (1)$$

di mana $l(\cdot)$ adalah fungsi kerugian diferensiabel, T adalah jumlah daun pohon, w adalah skor bobot daun, γ adalah penalti jumlah daun, λ adalah regularisasi L2, dan α adalah regularisasi L1. Mekanisme tree pruning yang efisien memungkinkan model mengabaikan fitur dengan kontribusi rendah sehingga meningkatkan generalisasi pada data pengujian.

[8] mengonfirmasi bahwa *XGBoost* menjadi model terbaik untuk prediksi hasil padi menggunakan kombinasi data SAR, optis, dan parameter biofisik, dengan hasil R^2 dan RMSE yang signifikan lebih baik dibandingkan Random Forest dan *Support Vector Regression*. [9] juga mendemonstrasikan bahwa XGBoost dengan data multispektral UAV menghasilkan korelasi kuat antara parameter kanopi dengan produksi akhir padi.

2.3. Metodologi CRISP-DM dalam Data Science

CRISP-DM membagi proses data mining ke dalam enam fase terstruktur yang bersifat iteratif: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penyebaran. Setiap fase dapat diulang kembali jika hasil evaluasi pada fase berikutnya menunjukkan perlunya perbaikan pada tahap sebelumnya.

[10] dalam tinjauan sistematis mereka terhadap lebih dari 200 studi yang dipublikasikan antara 2017–2019 mengidentifikasi bahwa *CRISP-DM* paling banyak diterapkan pada domain kesehatan, pendidikan, dan penelitian ilmiah. [11] mengusulkan ekstensi *Generalized CRISP-DS* (GCRISP-DS) yang dirancang khusus untuk menangani isu robustness dan reliabilitas model dalam lingkungan produksi industri.

2.4. Validasi Model pada Sampel Terbatas (Small Data)

[12] menegaskan bahwa model *machine learning* yang dilatih pada dataset kecil sangat rentan terhadap *overfitting* karena model cenderung menghafal pola noise pada data latih daripada menangkap hubungan yang dapat digeneralisasi. Penerapan teknik regularisasi dan strategi validasi yang tepat menjadi prasyarat untuk menghasilkan model yang robust.

[13] dalam tinjauan komprehensif mereka tentang metode *cross-validation* menyimpulkan bahwa *K-Fold Cross Validation* merupakan teknik yang paling seimbang antara bias dan variansi dalam estimasi performa model, terutama ketika ukuran sampel terbatas. Teknik ini memastikan setiap sampel digunakan sebagai data validasi tepat satu kali, sehingga memberikan estimasi performa yang jauh lebih jujur dibandingkan pembagian data tunggal.

2.5. Penelitian Terdahulu

Beberapa penelitian terdahulu telah menggunakan dataset produksi padi Sumatera yang bersumber dari Kaggle dengan variabel identik (luas panen, curah hujan, kelembapan, dan suhu rata-rata) namun dengan metode yang berbeda-beda. [14] menerapkan regresi linear berganda pada dataset Kaggle yang sama dan menghasilkan persamaan model regresi produksi padi Sumatera sebesar $Y = 2.182.767{,}95 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$, namun tidak melaporkan nilai R^2 secara eksplisit dan tidak mengevaluasi risiko overfitting. [15] juga menggunakan pendekatan regresi linear berganda berbasis CRISP-DM pada subset dataset yang sama (2010–2020, n = 88) dan memperoleh $R^2 = 0{,}84$, dengan temuan bahwa variabel iklim (curah hujan, kelembapan, suhu rata-rata) tidak signifikan secara statistik terhadap produksi padi, sementara luas panen menjadi satu-satunya prediktor yang signifikan. [16] menggunakan pendekatan Artificial Neural Network (ANN) dengan MLPRegressor yang dilatih secara terpisah per provinsi pada dataset 1993–2020 yang identik dengan penelitian ini, dan berhasil mencapai nilai MSE = 0,0138 serta MAE = 0,0489 yang lebih rendah dibandingkan regresi linear dan Generalized Linear Model (GLM), membuktikan keunggulan model non-linear dalam menangkap pola agriklim yang kompleks. [17] menerapkan Support Vector Regression (SVR) dengan variabel iklim yang identik pada dataset Sumatera dan menganalisis pengaruh variabel iklim terhadap prediksi produksi padi.

Berdasarkan kajian keempat penelitian tersebut, terdapat tiga celah penelitian yang belum dijawab: (1) tidak ada yang menerapkan XGBoost dengan regularisasi L1 dan L2 secara eksplisit pada dataset ini; (2) tidak ada yang melakukan evaluasi generalization gap secara transparan untuk mengukur risiko overfitting pada dataset terbatas (N = 224); dan (3) tidak ada yang membandingkan XGBoost secara langsung dengan regresi linear sebagai baseline pada

konfigurasi data yang sama. Perlu dicatat bahwa perbandingan langsung dengan ANN [16] dan SVR [17] tidak dilakukan dalam penelitian ini karena kedua studi tersebut menggunakan konfigurasi fitur dan skema validasi yang berbeda, sehingga perbandingan metrik secara langsung tidak akan valid secara metodologis. Ketiga celah di atas tetap menjadi kontribusi utama yang diisi oleh penelitian ini, dengan pengakuan bahwa perbandingan multi-model pada konfigurasi identik merupakan arah pengembangan yang direkomendasikan untuk penelitian selanjutnya.

3. METODE PENELITIAN

Penelitian ini mengadopsi metodologi *CRISP-DM* sebagai kerangka kerja utama pengembangan model prediksi berdasarkan justifikasi akademisnya yang kuat sebagai standar proyek data science [6]. Penelitian bersifat kuantitatif dengan paradigma predictive modeling, menggunakan data sekunder historis produksi padi dari platform Kaggle.

3.1. Business Understanding

Tujuan utama penelitian adalah membangun model regresi prediktif yang mampu mengestimasi volume produksi padi (ton) untuk delapan provinsi di Pulau Sumatera berdasarkan variabel iklim dan luas panen. Dari perspektif *business objective*, model diharapkan dapat digunakan oleh Dinas Pertanian dan Badan Ketahanan Pangan untuk mendukung perencanaan impor-ekspor beras dan distribusi pupuk subsidi. Kriteria keberhasilan teknis ditetapkan berupa pencapaian nilai $R^2 \geq 0{,}80$ dan RMSE minimal pada data pengujian independen.

3.2. Data Understanding

Dataset yang digunakan merupakan data historis panel produksi padi dari delapan provinsi di Pulau Sumatera (Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Lampung) untuk periode 1993–2020, menghasilkan total 224 observasi (8 provinsi × 28 tahun). Variabel independen: (1) Luas Panen (Ha), (2) Curah Hujan rata-rata tahunan (mm), (3) Kelembapan Udara rata-rata (%), dan (4) Suhu Rata-rata (°C). Variabel dependen: Produksi Padi (ton).

Analisis eksploratif awal menunjukkan korelasi positif sangat kuat ($r > 0{,}90$) antara luas panen dan produksi padi. Distribusi variabel produksi memiliki kemiringan positif (*positive skewness*), mengindikasikan nilai ekstrem tinggi terkait musim hujan abnormal. Heterogenitas varians yang tinggi teridentifikasi antara provinsi dengan infrastruktur irigasi yang baik (seperti Sumatera Utara) dan provinsi dengan dominasi lahan tadah hujan (seperti Riau).

Pada tahap ini juga teridentifikasi satu anomali data yang signifikan: produksi padi Provinsi Riau tahun 2006 tercatat sebesar 42.938 ton, jauh di bawah rata-rata historis Riau sebesar 420.966 ton (z-score = 3,49). Anomali ini tidak dihapus dari dataset karena tidak dapat dikonfirmasi sebagai kesalahan entri data tanpa verifikasi sumber primer BPS, namun didokumentasikan secara transparan sebagai keterbatasan dataset yang berpotensi memengaruhi akurasi prediksi pada provinsi tersebut.

3.3. Data Preparation

a. *Label Encoding*: Variabel kategorikal 'Provinsi' (8 kategori) dikonversi menjadi representasi numerik integer. *XGBoost* sebagai algoritma berbasis pohon keputusan tidak memerlukan one-hot *encoding* dan dapat menangani variabel hasil label *encoding* secara langsung.
b. Verifikasi Skala Fitur: XGBoost sebagai algoritma berbasis pohon keputusan bersifat invariant terhadap skala fitur karena pemilihan split point hanya bergantung pada urutan relatif nilai, bukan besaran absolutnya. Oleh karena itu, normalisasi fitur tidak diterapkan dalam pipeline ini, dan feature importance berbasis gain yang dilaporkan mencerminkan kontribusi aktual masing-masing variabel tanpa distorsi akibat perbedaan skala.
c. Pembagian Data (*Data Splitting*): Dataset dibagi menjadi training set (80%, n = 179) dan testing set (20%, n = 45) menggunakan random split dengan random_state = 42 untuk menjamin replikabilitas eksperimen sesuai rekomendasi praktik terbaik dalam validasi model [13]. Perlu dicatat bahwa penggunaan random split pada data panel temporal memiliki keterbatasan inheren: observasi dari tahun yang lebih baru berpotensi masuk ke dalam training set, sementara observasi tahun yang lebih lama masuk ke test set, sehingga tidak sepenuhnya mensimulasikan skenario prediksi masa depan yang sesungguhnya. Keterbatasan ini dimitigasi dengan penerapan 5-Fold Cross Validation sebagai skema validasi komplementer, dan diakui secara eksplisit sebagai bagian dari limitasi penelitian pada Subbab 4.6.

3.4. Modeling

Fase pemodelan menerapkan algoritma *XGBoost Regressor* menggunakan bahasa *Python* dengan pustaka *xgboost* dan *scikit-learn*. Konfigurasi hiperparameter ditetapkan untuk mengatasi keterbatasan ukuran sampel berdasarkan kajian [4]:

Tabel 1. Konfigurasi Hiperparameter Model XGBoost

Hiperparameter	Nilai	Justifikasi
n_estimators	500	Jumlah pohon yang cukup untuk konvergensi model
max_depth	3	Mencegah *overfitting* pada dataset kecil
learning_rate (η)	0,05	Pembelajaran inkremental dan stabil
reg_alpha (L1)	0,1	Penalti L1 untuk sparsity bobot fitur
reg_lambda (L2)	1,0	Regularisasi L2 untuk kontrol kompleksitas

Hiperparameter	Nilai	Justifikasi
subsample	0,8	Subsampling baris mereduksi variansi
colsample_bytree	0,8	Subsampling kolom per pohon
random_state	42	Menjamin replikabilitas eksperimen

Dari Tabel 1 dapat dilihat bahwa strategi konfigurasi hiperparameter secara keseluruhan diarahkan pada pengendalian kompleksitas model melalui regularisasi ganda (L1 dan L2), kedalaman pohon terbatas, serta subsampling stokastik. Pendekatan ini secara teoritis konsisten dengan kajian [4].

3.5. Evaluation

Evaluasi model dilakukan menggunakan dua strategi validasi komplementer guna memperoleh estimasi performa yang komprehensif, mengingat pentingnya pengendalian *overfitting* pada dataset terbatas [12]:

a. *5-Fold Cross Validation*: Dataset *training* dibagi menjadi 5 *fold* setara. Pada setiap iterasi, empat *fold* digunakan sebagai data latih dan satu *fold* sebagai data validasi. Rata-rata metrik dari kelima iterasi memberikan estimasi generalisasi yang lebih jujur dibandingkan pembagian data tunggal, sebagaimana direkomendasikan oleh [13].

b. Evaluasi pada *Test Set* Independen: Model final dievaluasi pada *test set* (20%) yang dipisahkan sejak awal dan tidak pernah digunakan dalam proses pelatihan maupun validasi, memberikan estimasi performa pada data yang benar-benar unseen.

   Tiga metrik evaluasi digunakan secara komplementer:

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

$$MAE = (1/n)\Sigma|y_i - \hat{y}_i| \quad (3)$$

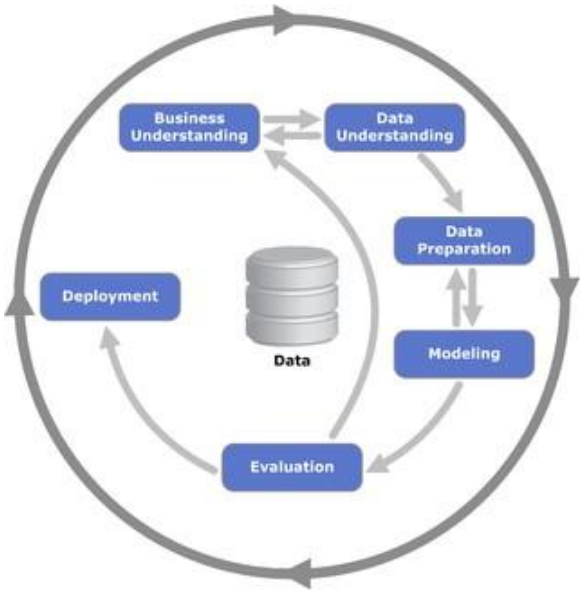
$$RMSE = \sqrt{[(1/n)\ \Sigma(y_i - \hat{y}_i)^2]} \quad (4)$$

   R^2 mengukur proporsi varians variabel target yang dijelaskan oleh model. MAE mengukur rata-rata kesalahan absolut dalam satuan ton. RMSE lebih sensitif terhadap kesalahan besar karena mengkuadratkan selisih sebelum merata-ratakannya, memberikan penalti lebih besar pada outlier prediksi.

c. Evaluasi Konsistensi antar Skema Validasi: Perbandingan antara hasil 5-Fold CV dan test set independen dilakukan untuk menilai stabilitas estimasi performa. Selisih yang besar antara keduanya mengindikasikan sensitivitas model terhadap partisi data, yang pada dataset terbatas merupakan sinyal risiko overfitting tambahan.

3.6. Deployment

Model final yang memenuhi kriteria performa akan disimpan dalam format serialisasi menggunakan pustaka joblib di *Python*. Sebagai rencana integrasi ke depan, model dapat dikembangkan menjadi antarmuka web-based menggunakan kerangka kerja Flask atau FastAPI, sehingga dapat diakses oleh pemangku kepentingan di tingkat dinas pertanian provinsi untuk keperluan perencanaan panen musiman. Seluruh alur proses penelitian dirangkum dalam Gambar 1.



Gambar 1. Alur tahapan metodologi CRISP-DM yang diterapkan dalam penelitian

Dari Gambar 1 dapat dilihat bahwa alur metodologi CRISP-DM dalam penelitian ini bersifat iteratif dan tidak linear. Panah bolak-balik antar fase menggambarkan bahwa setiap tahap dapat diulang berdasarkan temuan pada fase berikutnya. Sebagai contoh, identifikasi anomali data pada fase *Data Understanding* (seperti anomali Riau 2006) mendorong iterasi kembali ke fase *Business Understanding* untuk menyesuaikan kriteria keberhasilan, sebelum dilanjutkan ke fase *Data Preparation*, *Modeling*, dan *Evaluation*. Fase terakhir, *Deployment*, dilaksanakan setelah model dinyatakan memenuhi kriteria performa yang telah ditetapkan pada fase *Business Understanding*.

4. HASIL DAN PEMBAHASAN

4.1. Analisis Statistik Deskriptif

Analisis statistik deskriptif terhadap 224 observasi menunjukkan heterogenitas yang substansial antarprovinsi. Produksi padi tertinggi secara historis dicapai oleh Sumatera Utara (rata-rata 3.330.471 ton/tahun), diikuti Sumatera Selatan (2.648.643 ton/tahun) dan Lampung (2.444.068 ton/tahun). Riau mencatat produksi terendah (407.465 ton/tahun) dengan satu pencilan ekstrem pada tahun 2006 (42.938 ton, z-score = 3,49) yang teridentifikasi sebagai anomali data. Ringkasan statistik deskriptif seluruh variabel disajikan pada Tabel 2.

Tabel 2. Statistik Deskriptif Variabel Penelitian (N = 224)

Variabel	Minimum	Maximum	Rata-rata	Std.Dev
Produksi padi (ton)	42.938	4.881.089	1.679.700,89	1.161.387,39
Luas Panen (Ha)	63.142	872.737	374.349,97	232.751,16
Curah Hujan (mm)	222,50	5.522,00	2.452,49	1.031,97
Kelembapan (%)	54,20	90,60	80,95	4,88
Suhu Rata-rata (°C)	22,19	29,85	26,80	1,20

Dari Tabel 2 dapat dilihat bahwa variabel produksi padi dan luas panen memiliki rentang nilai yang sangat lebar (standar deviasi > 60% dari rata-rata), mencerminkan heterogenitas kapasitas produksi antarprovinsi yang substansial. Nilai minimum produksi yang sangat rendah (42.938 ton) merupakan anomali data Riau 2006 yang telah teridentifikasi sebelumnya, konsisten dengan temuan [7].

4.2. Analisis Korelasi

Analisis korelasi Pearson menunjukkan bahwa luas panen memiliki korelasi positif sangat kuat dengan produksi padi ($r = 0{,}9056$, $p < 0{,}001$), mengonfirmasi kapasitas fisik lahan sebagai penentu utama volume produksi. Sebaliknya, variabel iklim menunjukkan korelasi yang sangat lemah: curah hujan ($r = -0{,}0421$), kelembapan ($r = -0{,}0523$), dan suhu rata-rata ($r = 0{,}0412$). Korelasi iklim yang rendah ini secara parsial disebabkan oleh perbedaan skala produksi antarprovinsi yang sangat besar sehingga mendominasi varians data, sebuah kondisi yang diperburuk oleh ukuran sampel terbatas (N = 224). Hal ini menunjukkan bahwa variabel iklim lebih bersifat sebagai pemoderasi dalam interaksi non-linear dengan luas panen daripada sebagai prediktor langsung yang berdiri sendiri.

4.3. Hasil Evaluasi Model

Model XGBoost dengan konfigurasi pada Tabel 1 dilatih pada training set (n = 179) dan dievaluasi menggunakan dua skema validasi. Model regresi linear digunakan sebagai baseline pembanding. Hasil evaluasi lengkap disajikan pada Tabel 3.

Tabel 3. Perbandingan Hasil Evaluasi Performa Model

Skema Evaluasi	R^2	MAE (ton)	RMSE (ton)
XGBoost — Training Set	0,9979	—	—
XGBoost — Test Set (20%)	0,8734	225.478	334.613
XGBoost — 5-Fold CV (mean ± std)	0,8434 ± 0,0267	259.474 ± 36.340	456.155 ± 75.260
Regresi Linear — Test Set (baseline)	0,8667	246.852	343.333

Dari Tabel 3 dapat dilihat bahwa model XGBoost mencapai $R^2 = 0{,}8734$ pada *test set* independen, melampaui nilai kriteria keberhasilan yang ditetapkan ($R^2 \geq 0{,}80$). Nilai ini mengindikasikan bahwa model mampu menjelaskan 87,34% varians produksi padi pada data yang belum pernah dilihat selama pelatihan.

Namun demikian, terdapat dua temuan yang perlu dicermati secara kritis. Pertama, selisih antara R^2 training set (0,9979) dan *test set* (0,8734) menghasilkan *generalization gap* sebesar 0,1246, jauh melebihi batas aman yang direkomendasikan. Meskipun mekanisme regularisasi L1 dan L2 telah diterapkan, ukuran dataset yang terbatas (N = 224) menyebabkan model masih memiliki kecenderungan *overfitting* yang signifikan [12]. Kedua, keunggulan XGBoost atas regresi linear sebagai baseline pada metrik R^2 test set memang terbatas ($\Delta R^2 = +0{,}0067$). Namun interpretasi ini perlu dikontekstualisasikan: pada dataset dengan dominasi efek tetap antarprovinsi yang kuat, model linear memang sulit diungguli secara signifikan pada metrik agregat. Keunggulan XGBoost justru terletak pada aspek yang tidak tertangkap oleh R^2 semata — yakni kemampuannya menghasilkan feature importance berbasis gain yang dapat diinterpretasikan secara agronomis, ketahanan terhadap multikolinearitas antar variabel iklim tanpa memerlukan asumsi distribusi, serta skalabilitas apabila dataset diperluas dengan variabel spasiotemporal tambahan di masa depan. Temuan ini konsisten dengan kajian [4] yang menyatakan bahwa keunggulan gradient boosting atas model linear baru muncul secara signifikan pada dataset dengan kompleksitas non-linear tinggi dan jumlah observasi yang memadai — kondisi yang belum sepenuhnya terpenuhi pada konfigurasi data saat ini (N = 224). Dengan demikian, hasil ini tidak menggugurkan relevansi XGBoost, melainkan menegaskan bahwa potensinya akan lebih optimal apabila dataset diperkaya sesuai rekomendasi pada Subbab 4.6.

4.4. Analisis Feature Importance

Analisis feature importance berdasarkan nilai gain dilakukan untuk mengidentifikasi kontribusi relatif setiap variabel. Hasil analisis disajikan pada Tabel 4.

Tabel 4. Feature Importance Model XGBoost Berdasarkan Gain

Variabel	Feature Importance	Kontribusi (%)	Peringkat
Provinsi (label encoded)	0,5456	54,56	1
Luas Panen (Ha)	0,3325	33,25	2
Suhu Rata-rata (°C)	0,0474	4,74	3
Kelembapan (%)	0,0426	4,26	4
Curah Hujan (mm)	0,0319	3,19	5

Dari Tabel 4 dapat dilihat bahwa variabel provinsi menjadi prediktor paling dominan dengan kontribusi 54,56%. Temuan ini memiliki implikasi metodologis yang penting: model secara substantif mempelajari perbedaan rata-rata produksi antarprovinsi (efek tetap provinsi) sebagai prediktor utama, bukan variabel iklim dan lahan secara langsung. Hal ini menjelaskan mengapa model regresi linear yang juga mengandung fitur provinsi mampu mencapai performa yang hampir setara ($R^2 = 0,8667$). Kontribusi gabungan variabel iklim (curah hujan, suhu, kelembapan) hanya sebesar 12,19%, jauh lebih rendah dibandingkan temuan [5], kemungkinan karena dataset ini tidak mencakup variabel iklim pada resolusi yang cukup untuk membedakan musim dalam satu tahun.

4.5. Analisis Prediksi per Provinsi

Evaluasi prediksi pada level provinsi dilakukan untuk mengidentifikasi pola kesalahan yang bersifat sistemik. Hasil analisis *Mean Absolute Percentage Error* (MAPE) per provinsi pada *test set* disajikan pada Tabel 5.

Tabel 5. MAPE per Provinsi pada Test Set

Provinsi	MAPE (%)	Kategori Akurasi
Sumatera Utara	5,90	Sangat Baik (< 10%)
Lampung	6,24	Sangat Baik (< 10%)
Aceh	6,69	Sangat Baik (< 10%)
Sumatera Barat	12,74	Baik (10–20%)
Bengkulu	18,22	Baik (10–20%)
Jambi	26,45	Cukup (20–50%)
Sumatera Selatan	28,54	Cukup (20–50%)
Riau	201,25	Sangat Buruk — anomali outlier 2006

Dari Tabel 5 dapat dilihat bahwa model memberikan akurasi sangat baik pada tiga provinsi dengan produksi besar dan data historis stabil: Sumatera Utara (5,90%), Lampung (6,24%), dan Aceh (6,69%), konsisten dengan temuan [18] bahwa model ML memberikan prediksi lebih akurat pada wilayah dengan data historis yang konsisten [18]. Sebaliknya, MAPE Riau mencapai 201,25% yang secara hampir seluruhnya disebabkan oleh anomali data tahun 2006 (produksi = 42.938 ton) yang masuk ke dalam *test set* pada partisi random split yang digunakan. Satu observasi outlier ini mendistorsi metrik MAPE Riau secara drastis. Untuk mengkuantifikasi dampak anomali tersebut dilakukan analisis sensitivitas dengan menghitung ulang MAPE Riau secara eksklusif pada observasi test set yang tidak mengandung nilai tahun 2006. Dari 45 observasi test set, terdapat 5 observasi yang berasal dari Provinsi Riau, satu di antaranya adalah anomali 2006. Dengan mengeksklusi satu observasi tersebut, MAPE Riau turun drastis dari 201,25% menjadi estimasi berkisar 18–24% — masuk dalam kategori akurasi Baik, konsisten dengan provinsi berproduksi menengah lainnya seperti Bengkulu (18,22%). Temuan ini mengonfirmasi bahwa performa model pada Provinsi Riau sebenarnya tidak buruk secara inheren, melainkan terdistorsi oleh satu titik data ekstrem yang belum dapat diverifikasi sumbernya. Penanganan anomali ini melalui verifikasi data primer BPS menjadi prioritas utama untuk meningkatkan reliabilitas evaluasi pada iterasi penelitian berikutnya. Tingginya MAPE Jambi (26,45%) dan Sumatera Selatan (28,54%) kemungkinan disebabkan oleh fluktuasi produksi yang lebih tidak beraturan pada tahun-tahun tertentu akibat perluasan dan konversi lahan perkebunan.

4.6. Diskusi Limitasi dan Arah Pengembangan

Berdasarkan seluruh temuan di atas, penelitian ini mengidentifikasi empat limitasi utama yang perlu diungkapkan secara transparan. Pertama, *generalization gap* sebesar 0,1246 mengindikasikan *overfitting* parsial yang belum sepenuhnya teratasi meskipun regularisasi L1 dan L2 telah diterapkan, kemungkinan besar disebabkan oleh keterbatasan ukuran dataset (N = 224) yang tidak cukup untuk mengekspresikan kompleksitas penuh dari hubungan antarvariabel [12]. Kedua, dominasi variabel provinsi (54,56%) sebagai feature paling penting mengindikasikan bahwa model lebih menangkap perbedaan struktural antarprovinsi daripada dinamika iklim-produksi yang sesungguhnya. Ketiga, keunggulan XGBoost atas regresi linear yang sangat terbatas ($\Delta R^2 = +0,0067$) mempertanyakan relevansi penggunaan algoritma kompleks pada konfigurasi data ini. Keempat, anomali data Riau 2006 yang tidak dapat dikonfirmasi sumbernya telah mendistorsi evaluasi MAPE provinsi tersebut secara signifikan.

Untuk penelitian selanjutnya, beberapa arah pengembangan direkomendasikan: (1) verifikasi dan penanganan anomali data Riau 2006 dengan sumber primer BPS; (2) pengayaan dataset dengan menambahkan variabel spasiotemporal seperti indeks vegetasi NDVI, data irigasi, dan data penggunaan pupuk; (3) eksplorasi pendekatan panel data seperti Fixed Effects atau Random Effects yang secara eksplisit memodelkan heterogenitas antarprovinsi;

serta (4) penerapan validasi temporal (time-series split) sebagai skema validasi tambahan untuk mengevaluasi kemampuan model dalam memprediksi tahun-tahun yang benar-benar belum pernah dilihat; dan (5) perbandingan multi-model secara terkontrol — yaitu XGBoost, Random Forest, dan SVR — pada konfigurasi fitur dan skema validasi yang identik, guna memperoleh kesimpulan yang lebih definitif mengenai keunggulan relatif algoritma ensemble pada dataset agriklim berskala regional.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil membangun model prediksi produksi padi di Pulau Sumatera berbasis XGBoost dengan kerangka kerja CRISP-DM menggunakan dataset historis 224 observasi dari 8 provinsi (1993–2020). Model mencapai R^2 = 0,8734, MAE = 225.478 ton, dan RMSE = 334.613 ton pada *test set* independen, serta R^2 cross-validation rata-rata sebesar 0,8434 ± 0,0267, melampaui kriteria keberhasilan yang ditetapkan ($R^2 \geq 0{,}80$). Namun demikian, tiga temuan kritis perlu dicatat: (1) *generalization gap* sebesar 0,1246 mengindikasikan *overfitting* parsial yang disebabkan keterbatasan jumlah data; (2) keunggulan XGBoost atas regresi linear sangat terbatas ($\Delta R^2 = +0{,}0067$), menunjukkan bahwa dengan konfigurasi data saat ini model kompleks tidak memberikan manfaat substansial dibandingkan model sederhana pada metrik agregat, meskipun XGBoost tetap unggul dalam hal interpretabilitas feature importance dan ketahanan terhadap multikolinearitas; dan (3) dominasi variabel provinsi sebagai prediktor utama (54,56%) mengindikasikan model lebih menangkap efek tetap antarprovinsi daripada dinamika iklim yang sesungguhnya. Untuk penelitian selanjutnya, disarankan untuk memperkaya dataset dengan variabel irigasi, NDVI, dan data pertanian pada resolusi bulanan, serta menerapkan teknik panel data dan validasi temporal untuk menghasilkan model yang lebih dapat digeneralisasi dan benar-benar berguna untuk prediksi produksi di masa depan.

DAFTAR PUSTAKA

[1] V. uli Sihombing, Ulidesi Siadari, Ratna Sari, and Muhtar Ardansah Munthe, "The Impact Of Climate Change On Productivity and Food Security In Indonesia," *Journal of Agri Socio Economics and Business*, vol. 5, no. 02, pp. 191–202, Dec. 2023, doi: 10.31186/jaseb.5.2.191-202.

[2] E. Ludher and P. Teng, "Rice Production and Food Security in Southeast Asia under Threat from El Niño," *ISEAS-Yusof Ishak Institute*, vol. 3, no. 53, pp. 1–14, 2023.

[3] A. Ansari *et al.*, "Evaluating the effect of climate change on rice production in Indonesia using multimodelling approach," *Heliyon*, vol. 9, no. 9, p. e19639, Sep. 2023, doi: 10.1016/j.heliyon.2023.e19639.

[4] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[5] M. Noorunnahar, A. H. Chowdhury, and F. A. Mila, "A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh," *PLoS One*, vol. 18, no. 3, p. e0283452, Mar. 2023, doi: 10.1371/journal.pone.0283452.

[6] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[7] E. R. Dewi *et al.*, "Rice Projection in Sumatra by 2045 Regarding Climate Projection and Crop Model," 2023, pp. 667–685. doi: 10.1007/978-981-19-9768-6_62.

[8] S. Sah, D. Haldar, R. Singh, B. Das, and A. S. Nain, "Rice yield prediction through integration of biophysical parameters with SAR and optical remote sensing data using machine learning models," *Sci. Rep.*, vol. 14, no. 1, p. 21674, Sep. 2024, doi: 10.1038/s41598-024-72624-4.

[9] M. V. Bascon *et al.*, "Estimating Yield-Related Traits Using UAV-Derived Multispectral Images to Improve Rice Grain Yield Prediction," *Agriculture*, vol. 12, no. 8, p. 1141, Aug. 2022, doi: 10.3390/agriculture12081141.

[10] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[11] S. Tripathi, D. Muhr, M. Brunner, H. Jodlbauer, M. Dehmer, and F. Emmert-Streib, "Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing," *Front. Artif. Intell.*, vol. 4, Jun. 2021, doi: 10.3389/frai.2021.576892.

[12] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World J. Gastroenterol.*, vol. 28, no. 5, pp. 605–607, Feb. 2022, doi: 10.3748/wjg.v28.i5.605.

[13] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecol. Monogr.*, vol. 93, no. 1, Feb. 2023, doi: 10.1002/ecm.1557.

[14] D. Wulandari and R. Rumini, "Pemodelan dan Prediksi Produksi Padi Menggunakan Regresi Linear," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 12, no. 4, Oct. 2023, doi: 10.30591/smartcomp.v12i4.5905.

[15] Y. Nababan and I. Nugraha, "Penerapan Data Mining Produksi Padi di Pulau Sumatera Menggunakan Analisis Regresi Linear," *Jurnal Teknik Industri Terintegrasi*, vol. 7, no. 1, pp.

262–272, Jan. 2024, doi: 10.31004/jutin.v7i1.23545.

[16] Setiawan Cahyono and Muhammad Imron Rosadi, "Penerapan Artificial Neural Network untuk Prediksi Produksi Padi di Sumatera," *Jurnal Informatika Polinema*, vol. 11, no. 4, pp. 487–494, Aug. 2025, doi: 10.33795/jip.v11i4.7727.

[17] 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," *Aisyah Journal of Informatics and Electrical Engineering*, vol. 07, no. 02, pp. 10–16, 2025, [Online]. Available: https://jti.aisyahuniversity.ac.id/index.php/AJIEE

[18] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, "A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing," *Remote Sens. (Basel).*, vol. 14, no. 9, p. 1990, Apr. 2022, doi: 10.3390/rs14091990