Journal of Computer Science and Computational Intelligence
May 2025 | Vol. 1 | No.1 | Page 9-20
E-ISSN: 3109-7383 P-ISSN: 3110-8393
DOI: https://doi.org/10.30598/algorithm.v1i1.9-20

Prediksi Produktivitas Padi Menggunakan Algoritma Random Forest di Provinsi Sumatera Tahun 1993–2020

(Prediction of Rice Productivity Using the Random Forest Algorithm in Sumatera Province from 1993-2020)

Putri Aprilia de Feretes[1*], Shinta Rante Mangaluk[2]

[1,2]Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Pattimura
Jl. Ir. M. Putuhena, Ambon, 97233, Indonesia
[*]Corresponding author's e-mail: * **putridefretes@gmail.com**

Manuscript submitted: ***14th February 2025*** | ***Manuscript revision:*** ***27th March 2025*** | ***Accepted for publication:*** ***29th May 2025***

Abstract

This study aims to analyze the relationship between rice production and environmental factors in the Sumatra region using data science approaches and machine learning algorithms. The dataset used includes information on rice production, harvest area, rainfall, humidity, and average temperature from various provinces in Sumatera between 1993 and 2020. The analysis was conducted through data exploration, Pearson correlation test, feature engineering such as environmental index and annual temperature fluctuation, and predictive model building using linear regression, decision tree, and Random Forest algorithms. The results showed that harvest area had the highest correlation to rice production, while environmental factors also showed significant influence. The Random Forest model was selected as the best model based on the evaluation of R^2, mean average error, and root mean square error metrics. In addition, parameter tuning and cross-validation were conducted to improve model performance. This study emphasizes the importance of utilizing data-driven quantitative approaches in supporting more precise agricultural planning and policies.

Keywords: *Rice Paddy Production; Data Science; Machine Learning; Random Forest; Harvest Area; Environmental Factors; Regression; Prediction; Sumatera; Precision Agriculture.*

How to cite this article:
P. de Fretes and S. Mangaluk, "Prediksi Produktivitas Padi Menggunakan Algoritma Random Forest di Provinsi Sumatera Tahun 1993–2020", *algorithm*, vol. 1, no. 1, pp. 9-20, May 2025.

Journal homepage: https://ojs3.unpatti.ac.id/index.php/algorithm
Research Article Open Access

1. PENDAHULUAN

Sektor pertanian memegang peranan penting dalam pembangunan ekonomi Indonesia, terutama di wilayah Sumatera yang memiliki potensi besar dalam produksi tanaman pangan, khususnya padi. Produksi padi tidak hanya dipengaruhi oleh faktor manusia dan kebijakan, tetapi juga oleh faktor lingkungan seperti curah hujan, kelembaban udara, dan suhu rata-rata [1]. Oleh karena itu, pemanfaatan data historis terkait produksi padi dan faktor-faktor lingkungan sangat penting untuk mendukung pengambilan keputusan yang lebih tepat. Dengan berkembangnya ilmu data dan teknologi kecerdasan buatan, kini memungkinkan untuk menganalisis hubungan antara berbagai variabel yang mempengaruhi produksi padi serta membangun model prediktif yang dapat membantu perencanaan pertanian yang lebih baik [2]. Indonesia merupakan negara agraris yang sebagian besar penduduknya bergantung pada sektor pertanian sebagai mata pencaharian utama. Salah satu komoditas strategis yang menjadi tulang punggung ketahanan pangan nasional adalah padi. Padi tidak hanya menjadi sumber utama karbohidrat bagi masyarakat Indonesia, tetapi juga berperan penting dalam menjaga stabilitas ekonomi dan sosial [3]. Wilayah Sumatera, sebagai salah satu pulau besar di Indonesia, memiliki peran signifikan dalam menyumbang produksi padi nasional. Tiap provinsi di Sumatera memiliki karakteristik geografis dan iklim yang unik, yang berdampak langsung terhadap hasil produksi pertanian. Oleh karena itu, penting untuk memahami faktor-faktor yang mempengaruhi produktivitas tanaman padi di wilayah ini [4]. Produksi padi sangat dipengaruhi oleh beberapa faktor, baik yang bersifat teknis seperti luas panen dan jenis varietas, maupun faktor lingkungan seperti curah hujan, kelembaban udara, dan suhu rata-rata.

Faktor krusial dalam menjaga ketahanan pangan adalah memperluas lahan pertanian serta menghindari alih fungsi lahan pertanian menjadi lahan nonpertanian. [5]. Berangkat dari latar belakang tersebut, proyek ini bertujuan untuk menganalisis data produksi padi di wilayah Sumatera dari tahun ke tahun, dengan mempertimbangkan pengaruh variabel lingkungan dan luas panen. Melalui analisis korelasi dan pemodelan prediktif menggunakan algoritma *Machine learning*, diharapkan hasil dari proyek ini dapat memberikan kontribusi nyata dalam pengembangan pertanian presisi dan pengambilan keputusan yang lebih akurat.

Produktivitas padi merupakan indikator penting dalam mengukur keberhasilan sektor pertanian suatu wilayah. Seiring meningkatnya kebutuhan pangan nasional dan dinamika perubahan iklim, analisis terhadap faktor-faktor yang mempengaruhi produksi padi menjadi semakin penting. Sisi positif dan negatif curah hujan bagi pertanian padi sawah adalah bersifat dua mata pisau di satu sisi ia menawarkan potensi besar untuk hasil panen yang sukses, namun di sisi lain, ia juga dapat menjadi ancaman serius terhadap kemandirian pangan. [6]. Sejalan dengan itu, perkembangan teknologi analitik seperti *machine learning* memberikan pendekatan baru dalam pemodelan prediksi hasil panen. Namun, upaya prediksi canggih yang berusaha memahami keterkaitan antara variabel ini tetap berhadapan dengan masalah mendasar di lapangan, sebab pengelolaan sumber daya air yang buruk tidak hanya berdampak negatif pada kesehatan masyarakat, tetapi juga secara langsung mengurangi produktivitas pertanian—sebuah sektor yang menjadi tulang punggung ekonomi banyak negara berkembang. [7].

Untuk memantau kualitas dan membuat perkiraan hasil panen, metode machine learning dan computer vision digunakan dalam rangka mengklasifikasikan berbagai data visual (set gambar) dari tanaman [8]. Bukan hanya untuk itu klasifikasi *Random Forest* juga dapat dimanfaatkan untuk pengukuran tingkat penyakit pada tumbuhan sehingga dapat mempengaruhi hasil panen [9]. Penelitian ini memberikan inspirasi dalam merancang fitur-fitur turunan seperti indeks lingkungan dan fluktuasi suhu tahunan. Pendekatan regresi dan algoritma *Random Forest* terbukti efektif dalam berbagai studi. Salah satu penelitian, misalnya, mengembangkan model prediksi produktivitas padi di Kabupaten Batang dengan data citra Sentinel-2 dan menghasilkan akurasi klasifikasi tutupan lahan hingga 91%, serta RMSE prediksi berkisar antara 0,498 hingga 1,857

ton/ha tergantung pada data validasi yang digunakan [10]. Model *Random Forest* terbukti mampu menangani kompleksitas data dan menghasilkan prediksi yang cukup akurat untuk pengambilan keputusan berbasis data (*data-driven policy*) di bidang pertanian.

2. METODE PENELITIAN

2.1. Pemrosesan Data

Dataset yang digunakan dalam penelitian ini terdiri dari 224 entri data yang mencakup beberapa provinsi di wilayah Sumatera dalam rentang waktu dari tahun 1993 hingga 2020. Fitur-fitur yang tersedia dalam dataset ini meliputi tujuh variabel utama. Pertama, *Provinsi* yang menunjukkan nama provinsi tempat pengambilan data. Kedua, *Tahun* yang merepresentasikan tahun pengamatan. Ketiga, *Produksi (ton)* yang merupakan jumlah produksi padi dalam satuan ton. Keempat, *Luas Panen (hektar)* yang menunjukkan luas lahan yang digunakan untuk penanaman padi. Kelima, *Curah Hujan (mm)* yang mencerminkan total curah hujan selama periode waktu tertentu. Keenam, *Kelembaban (%)* yang menunjukkan tingkat kelembaban rata-rata, dan ketujuh adalah *Suhu Rata-rata (°C)* yang menggambarkan suhu rata-rata selama periode pengamatan tersebut.

Langkah-langkah pemrosesan data dilakukan secara sistematis untuk memastikan kualitas dan kesiapan data sebelum digunakan dalam pelatihan model. Tahap pertama adalah pemeriksaan nilai kosong, dengan tujuan memastikan bahwa tidak terdapat nilai yang hilang (*missing values*) dalam dataset. Tahap kedua adalah standarisasi nama kolom, yang dilakukan untuk mempermudah proses analisis dan menjaga konsistensi penamaan. Selanjutnya, dilakukan visualisasi distribusi data serta deteksi outlier menggunakan diagram Boxplot untuk mengidentifikasi nilai-nilai pencilan yang dapat memengaruhi kualitas model. Tahap terakhir adalah normalisasi data, di mana fitur-fitur numerik distandarisasi menggunakan *StandardScaler* agar semua fitur memiliki skala yang sama, sehingga menghindari dominasi salah satu fitur terhadap model.

2.2. Analisis Korelasi

Analisis korelasi digunakan untuk mengidentifikasi hubungan antara variabel independen (fitur) dengan variabel dependen (produksi padi). Korelasi Pearson digunakan untuk mengukur kekuatan hubungan linier antara variabel:

1. Luas Panen: Diharapkan memiliki korelasi tinggi terhadap produksi padi karena semakin besar luas lahan, semakin tinggi potensi produksi.
2. curah hujan, Kelembaban, dan Suhu Rata-rata: Faktor-faktor lingkungan ini akan diperiksa kontribusinya secara statistik terhadap produksi padi. Hasil korelasi membantu dalam memilih fitur yang paling relevan dalam pembangunan model prediksi.

2.3. Pembuatan Data *(Feature Engineering)*

Dalam tahap ini, dilakukan rekayasa fitur untuk memperkaya model prediksi. Beberapa fitur yang digunakan antara lain: pertama, rata-rata curah hujan per tahun per provinsi yang merupakan agregasi tahunan untuk melihat pola curah hujan sepanjang tahun. Kedua, perubahan suhu dari tahun ke tahun yang dianalisis untuk mengamati fluktuasi suhu sebagai faktor yang dapat memengaruhi pertumbuhan padi. Ketiga, faktor indeks lingkungan yang merupakan gabungan dari data curah hujan, suhu, dan kelembaban untuk melihat dampaknya secara komprehensif. Pendekatan ini mengacu pada sistem air-lingkungan–tanaman yang dikembangkan oleh Alfarisy et al. (2024). Setelah proses rekayasa fitur selesai, dataset kemudian dibagi menjadi dua bagian, yaitu fitur (X) dan target (y), dengan target berupa produksi padi.

2.4. Pembuatan Model Prediksi

Model prediksi dikembangkan menggunakan algoritma regresi dengan tiga pilihan model

utama, yaitu Linear Regression, Random Forest Regressor, dan Decision Tree Regressor. Linear Regression digunakan sebagai model dasar untuk analisis hubungan linier antar variabel. Random Forest Regressor, sebagai model ensemble, dipilih karena kemampuannya yang lebih kuat dalam menangani data dengan variabilitas tinggi. Sementara itu, Decision Tree Regressor digunakan karena sifatnya yang non-linear dan kemudahannya dalam interpretasi hasil.

Proses pelatihan dan evaluasi model dilakukan melalui beberapa tahapan. Pertama, dataset dibagi menjadi dua bagian, yaitu data latih sebanyak 80% dan data uji sebanyak 20%. Setelah itu, model dilatih menggunakan data latih untuk memahami hubungan antara fitur-fitur input dan target output. Evaluasi terhadap performa model dilakukan menggunakan metrik seperti koefisien determinasi (R^2), Mean Absolute Error (MAE), dan Root Mean Square Error (RMSE) untuk mengukur seberapa baik model dalam memprediksi produksi padi pada data uji. Model terbaik kemudian dipilih berdasarkan skor evaluasi tertinggi yang diperoleh dari hasil pengujian beberapa model.

Dataset dibagi ke dalam dua bagian utama, yaitu data pelatihan dan data pengujian, dengan proporsi sebesar 80:20. Model kemudian dilatih selama sejumlah epoch, di mana proses pelatihan disertai dengan validasi secara simultan guna meminimalkan risiko terjadinya overfitting pada model. Berdasarkan hasil eksperimen yang telah dilakukan, diperoleh bahwa konfigurasi model terbaik dicapai ketika menggunakan 128 neuron dengan nilai learning rate sebesar 0.01. Pada konfigurasi tersebut, model berhasil menghasilkan nilai Mean Squared Error (MSE) sebesar 232.283.171.177,98.

2.5. Pembuatan dan Training Model

Model prediksi dibangun menggunakan beberapa algoritma regresi, yaitu Linear Regression, Random Forest Regressor, dan Decision Tree Regressor. Linear Regression digunakan sebagai model dasar untuk menganalisis hubungan linier antar variabel. Random Forest Regressor dipilih sebagai model ensemble karena lebih kuat dalam menangani data dengan variabilitas tinggi. Sementara itu, Decision Tree Regressor digunakan sebagai model non-linear yang mudah diinterpretasi.

Langkah-langkah yang dilakukan dalam pembangunan model terdiri dari empat tahap utama. Pertama, proses pembagian data dilakukan dengan membagi dataset menjadi data latih (80%) dan data uji (20%) menggunakan metode train-test split. Kedua, model dilatih menggunakan data latih untuk mempelajari hubungan antara fitur-fitur input dengan target output. Ketiga, evaluasi model dilakukan dengan menggunakan metrik seperti koefisien determinasi (R^2), Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE) untuk mengukur seberapa baik model dalam memprediksi produksi padi pada data uji. Terakhir, model terbaik dipilih berdasarkan skor evaluasi tertinggi yang diperoleh dari sejumlah model yang telah diuji, dan model tersebut akan digunakan untuk melakukan prediksi.

2.6. *Hyperparameter tuning* dan *Cross-Validation*

Untuk meningkatkan performa model, dilakukan pencarian hyperparameter menggunakan GridSearchCV. Beberapa parameter yang di-tuning termasuk jumlah estimator (n_estimators), kedalaman maksimum pohon keputusan (max_depth), dan jumlah sampel minimum untuk pemisahan (min_samples_Split). Proses tuning ini bertujuan untuk mencari kombinasi parameter terbaik dari model *Random Forest*. Selain itu, *cross-validation* (5-*Fold*) digunakan untuk memeriksa performa model secara lebih *robust* di berbagai subset data, sehingga dapat menghindari masalah *overfitting* dan memastikan generalisasi yang lebih baik pada data yang belum terlihat.

2.7. Penyimpanan Model dan *Scaler*

Setelah model dilatih dan diuji, model dan Scaler yang digunakan untuk normalisasi fitur disimpan menggunakan joblib untuk dapat digunakan kembali di masa mendatang dalam proses prediksi.

2.8. Prediksi Data Baru

Model yang telah dilatih digunakan untuk memprediksi produksi padi berdasarkan data Input baru yang diberikan, seperti luas panen, curah hujan, kelembaban, dan suhu rata-rata. Model memberikan estimasi produksi padi untuk data baru.

3. HASIL DAN PEMBAHASAN

3.1. Pemrosesan Data

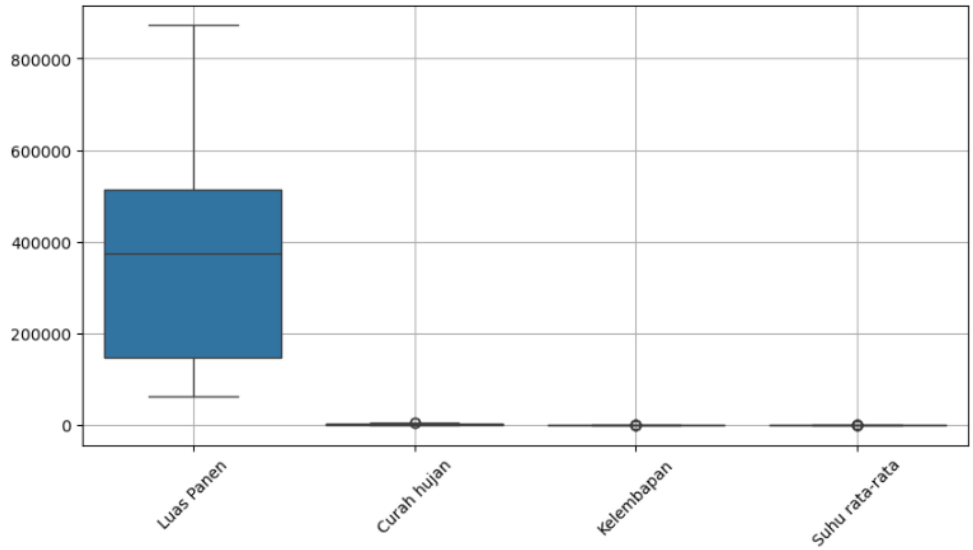
Dataset yang digunakan dalam penelitian ini berisi 224 entri data yang mencakup beberapa provinsi di Pulau Sumatera, dengan rentang waktu dari tahun 1993 hingga 2020. Setiap entri memuat informasi terkait Produksi Padi, Luas Panen, curah hujan, Kelembaban, dan Suhu Rata-rata. Langkah awal dalam pemrosesan data adalah memastikan bahwa dataset tidak mengandung nilai kosong *(missing values)*. Berdasarkan hasil pemeriksaan, dataset dinyatakan lengkap dan siap digunakan untuk tahap analisis berikutnya. Untuk menjaga konsistensi dan keterbacaan data, dilakukan standarisasi nama kolom sehingga seluruh nama variabel menjadi seragam dan lebih mudah dipahami. Selanjutnya, dilakukan deteksi keberadaan *Outlier* melalui visualisasi distribusi data menggunakan Boxplot. Hasil visualisasi menunjukkan bahwa terdapat beberapa *Outlier*, khususnya pada fitur Luas Panen dan curah hujan. Karena keberadaan *Outlier* dapat berpotensi mempengaruhi kinerja model prediktif, maka diambil langkah normalisasi data untuk menyamakan skala antar fitur. Normalisasi dilakukan dengan menggunakan metode StandardScaler, yang mengubah distribusi data sehingga memiliki nilai rata-rata (mean) sebesar 0 dan standar deviasi sebesar 1, guna mencegah dominasi fitur tertentu dalam proses pembelajaran model secara rinci, tahapan preprocessing yang dilakukan dapat dijelaskan sebagai berikut:

1. *Load Dataset*. Pada bagian ini, berbagai library yang diperlukan untuk analisis dipanggil, seperti pandas untuk manipulasi data, numpy untuk operasi numerik, matplotlib dan seaborn untuk visualisasi, serta library dari sklearn untuk pemodelan dan evaluasi. Dataset kemudian dimuat ke dalam sebuah DataFrame df menggunakan pd.read_csv.
2. Pisahkan Fitur dan Target Setelah dataset dimuat, dilakukan pemisahan antara fitur (X) dan target (y). Fitur yang digunakan adalah variabel-variabel Input (Luas Panen, curah hujan, Kelembaban, Suhu rata-rata), sedangkan targetnya adalah Produksi padi.
3. Normalisasi Fitur Numerik Untuk memastikan semua fitur berada pada skala yang sama, dilakukan normalisasi menggunakan StandardScaler dari sklearn. StandardScaler bekerja dengan cara mengubah data sehingga memiliki $rata - rata\ (mean) = 0$ $dan\ standar\ deviasi = 1$. Ini penting untuk mencegah fitur dengan nilai besar (seperti curah hujan) mendominasi pembelajaran model.
4. *Split* data Setelah normalisasi, data dibagi menjadi data latih (train) dan data uji (test) dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian (test_size=0.2). Parameter $random\_state = 42$ digunakan agar hasil pembagian data dapat direproduksi (konsisten saat dijalankan ulang).
5. Deteksi *Outlier* dengan Boxplot pada bagian ini, dilakukan visualisasi distribusi data dengan Boxplot menggunakan seaborn. Boxplot ini berguna untuk mendeteksi *outlier* pada

fitur numerik. *Outlier* biasanya terlihat sebagai titik-titik di luar "kumis" (whisker) pada Boxplot. Adanya *outlier* terdeteksi terutama pada luas panen dan curah hujan, yang selanjutnya menjadi perhatian khusus dalam tahap preprocessing data.

3.2. Analisis Korelasi

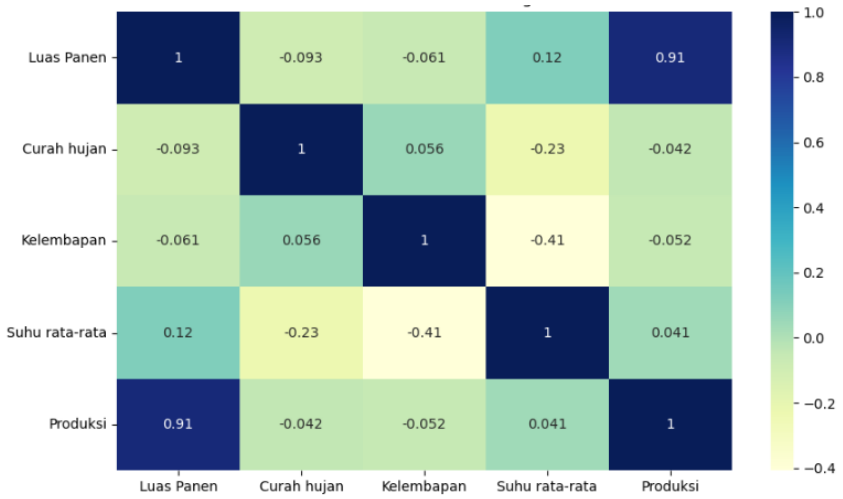
Analisis korelasi antar fitur menunjukkan bahwa variabel store memiliki korelasi negatif sedang terhadap weekly_sales dengan nilai korelasi sebesar -0.32, sedangkan fitur temperature, cpi, dan unemployment menunjukkan hubungan yang tergolong lemah terhadap weekly_sales, dan korelasi tertinggi antar fitur ditemukan antara temperature dan fuel_price sebesar 0.12 serta antara cpi dan *temperature* sebesar 0.22, yang seluruhnya divisualisasikan melalui matriks korelasi, dapat dilihat pada Gambar 1.



Gambar 1. Deteksi *Outlier* Pada Fitur Numerik Menggunakan Boxplot.

3.3. Analisis Korelasi

Analisis korelasi dilakukan untuk memahami hubungan antar variabel numerik terhadap produksi padi. Dalam hal ini, digunakan korelasi Pearson untuk mengukur kekuatan hubungan linier antara variabel-variabel yang diamati. Hasil analisis menunjukkan bahwa terdapat korelasi positif yang signifikan antara luas panen dan produksi padi, yang mengindikasikan bahwa semakin besar luas lahan yang digunakan, maka semakin tinggi potensi produksi padi. Selain Luas Panen, faktor lingkungan seperti curah hujan, Kelembaban, dan Suhu Rata-rata juga diperiksa kontribusinya secara statistik terhadap produksi padi. Hasil analisis menunjukkan bahwa variabel-variabel lingkungan tersebut memiliki korelasi moderat terhadap produksi. Meskipun tidak sekuat pengaruh luas panen, faktor-faktor ini tetap memberikan kontribusi yang relevan dalam mendukung hasil panen. Temuan ini menegaskan bahwa, selain faktor lahan, kondisi lingkungan juga memegang peran penting dalam menentukan tingkat produksi padi.



Gambar 2: Korelasi Antara Fitur-Fitur Dan Target Produksi Padi.

Berdasarkan grafik yang ditampilkan, dapat diamati bahwa kurva *training loss* maupun validation loss mengalami penurunan secara bertahap dan konsisten, yang menandakan proses pelatihan berjalan dengan stabil. Kedua kurva mulai mencapai titik konvergensi sekitar *epoch* ke-50, yang menunjukkan bahwa model telah berhasil menemukan pola data yang relevan tanpa menunjukkan gejala *overfitting*. Ketiadaan jarak signifikan antara kurva *training* dan *validation* memperkuat indikasi bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya, serta mampu melakukan prediksi secara akurat dan efisien di luar data pelatihan.

3.4. Pembuatan Data *(Feature Engineering)*

Dalam tahap *feature engineering*, jika diperlukan, beberapa fitur baru dibuat untuk memperkaya model maupun analisis. Adapun fitur-fitur baru yang dibuat antara lain:

1. Rata-rata curah hujan per tahun per provinsi: untuk melihat pola curah hujan sepanjang tahun dan apakah terdapat tren yang mempengaruhi produksi padi.
2. Perubahan Suhu dari Tahun ke Tahun: untuk melihat fluktuasi suhu yang dapat mempengaruhi pertumbuhan tanaman padi.
3. Indeks Lingkungan: gabungan dari variabel curah hujan, suhu, dan kelembaban yang bertujuan untuk mengukur dampak gabungan faktor-faktor lingkungan terhadap produksi padi. Penggunaan indeks lingkungan mengacu pada penelitian yang dilakukan oleh [3], yang menunjukkan bahwa faktor-faktor lingkungan saling terkait dalam mendukung hasil pertanian.

Setelah pembuatan fitur tambahan, data dibagi menjadi dua bagian: fitur (X) yang terdiri dari variabel Luas Panen, curah hujan, Kelembaban, Suhu Rata-rata, serta fitur tambahan; dan target (y) yang berupa Produksi padi.

3.5. Training Model

Tiga model regresi diuji untuk memprediksi Produksi padi:

1. *Linear regression*, yang digunakan sebagai model dasar untuk menganalisis hubungan linier antara fitur dan target.

2. *Random Forest regressor*, yang lebih kompleks dan kuat, cocok untuk menangani data dengan variabilitas tinggi.
3. *Decision Tree Regressor*, model non-linear yang memberikan interpretasi lebih mudah tentang keputusan yang diambil oleh model.

Model-model ini dilatih menggunakan 80% data latih dan diuji pada 20% data uji. Evaluasi dilakukan menggunakan tiga metrik: R^2 (Koefisien Determinasi), MAE (Mean Absolute Error), dan RMSE *(Root Mean Squared Error)*. Hasil evaluasi menunjukkan bahwa *Random Forest* Regressor memberikan performa terbaik, dengan skor R^2 yang lebih tinggi dibandingkan dengan model *linear* regression dan *decision tree*. Model ini juga memiliki RMSE yang lebih rendah, yang menandakan bahwa prediksi model ini lebih akurat. Secara rinci, tahapan yang dilakukan dapat dijelaskan sebagai berikut:

1. Pembuatan Model *Random Forest*
   Pada tahap ini, dibuat sebuah model prediksi menggunakan algoritma *Random Forest regressor*. *Random Forest* Regressor adalah metode *ensemble* berbasis pohon keputusan yang bekerja dengan membangun banyak pohon keputusan pada data latih dan mengeluarkan rata-rata prediksi dari masing-masing pohon untuk meningkatkan akurasi dan mengontrol *overfitting*. Parameter $n\_estimators = 100$ berarti model akan membangun 100 pohon dalam hutan. Parameter $random\_state = 42$ digunakan untuk memastikan hasil yang konsisten setiap kali kode dijalankan, karena mengatur seed randomisasi.

2. Melatih Model
   Setelah model *Random Forest* dibuat, langkah selanjutnya adalah melakukan proses pelatihan (training) dengan data latih. fit(X_train, y_train) berarti model dilatih menggunakan fitur Input (X_train) dan target Output (y_train) dari dataset latih. Pada tahap ini, *Random Forest* membangun dan menyesuaikan 100 pohon keputusan berdasarkan data pelatihan, sehingga model dapat belajar hubungan antara fitur Input dan target Output untuk prediksi di masa mendatang.

3. Evaluasi Model

Tabel 1. Hasil Evaluasi Model Pada Data Uji

No	Metode Evaluasi	Nilai
1.	*Root Mean Squared Error (RMSE)*	322494.34
2.	*Mean Absolute Errror (MAE)*	216003.18
3.	*R^2 Score (Test Set)*	0.88

Berdasarkan hasil evaluasi model yang ditunjukkan pada Tabel 1, performa model pada data uji dinilai menggunakan tiga metrik utama, yaitu *Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)*, dan R^2 *Score*. Nilai *RMSE* yang diperoleh sebesar 322.494,34 menunjukkan rata-rata besarnya deviasi prediksi terhadap nilai aktual, sedangkan nilai *MAE* sebesar 216.003,18 merepresentasikan rata-rata kesalahan absolut yang lebih mudah ditafsirkan karena tidak dikuadratkan. Sementara itu, nilai R^2 Score pada data uji mencapai 0,88, yang berarti model mampu menjelaskan sekitar 88% variasi dari data aktual, sehingga dapat dikatakan bahwa model memiliki performa yang cukup baik pada data uji.

Namun, hasil *cross-validation* yang ditampilkan pada Tabel 2 menunjukkan nilai R^2 Score rata-rata sebesar -1,4581 ± 4,0399. Nilai negatif ini mengindikasikan bahwa model tidak mampu melakukan generalisasi dengan baik ketika divalidasi menggunakan metode *k-fold cross-validation (5-Fold)*, bahkan performanya lebih buruk dibandingkan prediksi sederhana berupa rata-rata nilai target. Selain itu, standar deviasi yang besar menunjukkan adanya ketidakstabilan model ketika diuji pada berbagai lipatan data. Dengan demikian, meskipun model terlihat baik pada data uji dengan R^2 tinggi, hasil validasi silang mengindikasikan adanya potensi overfitting, sehingga diperlukan perbaikan baik dari sisi pemilihan fitur, parameter model, maupun strategi validasi agar performa model lebih konsisten.

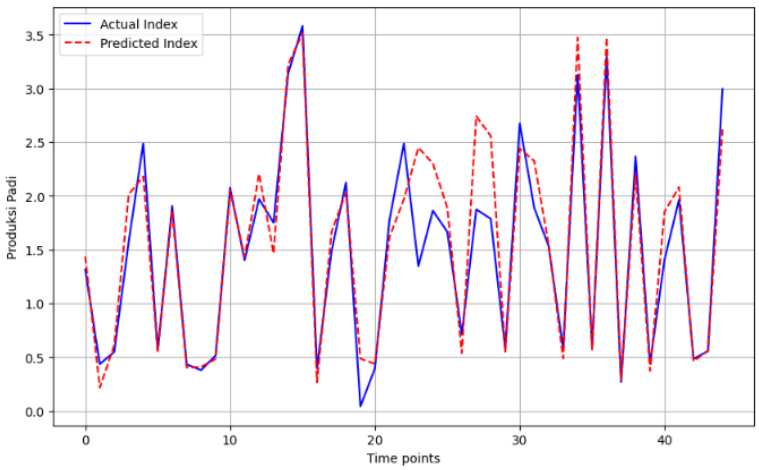
Tabel 2. Hasil *Cross-Validation* (5-*Fold*)

No	Metode Evaluasi	Nilai
1.	*R^2 Score Cross-Validation (5-Fold)*	-1.4581 ± 4.0399

4. Visualisasi hasil prediksi (dengan grafik garis)

Visualisasi hasil prediksi dilakukan untuk membandingkan nilai aktual produksi dengan hasil estimasi model. Perbandingan tersebut disajikan pada **Gambar 3**, yang memperlihatkan hubungan antara data aktual (garis biru) dan hasil prediksi (garis merah). Secara umum, pola kedua kurva menunjukkan kecenderungan yang serupa, menandakan bahwa model memiliki kemampuan yang baik dalam menangkap pola tren dari data historis produksi. Kesamaan fluktuasi antar garis menunjukkan bahwa model mampu mempelajari dinamika variabel input secara efektif sehingga menghasilkan nilai prediksi yang mendekati kondisi aktual.

Visualisasi seperti ini penting dilakukan untuk mengidentifikasi sejauh mana model mampu melakukan *generalization* terhadap data yang belum pernah dilihat sebelumnya. Pola kesesuaian antara hasil prediksi dan data aktual menjadi indikator awal kinerja model sebelum dilakukan evaluasi kuantitatif menggunakan metrik seperti MAE, RMSE, dan R^2. Hasil yang stabil dan konsisten antara prediksi dan aktual menunjukkan bahwa model tidak mengalami *overfitting* maupun *underfitting.*

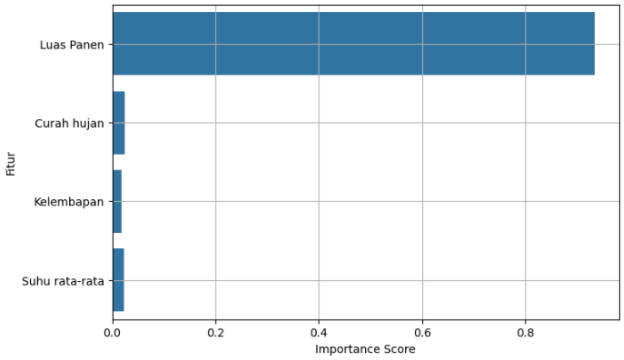


Gambar 3. Perbandingan hasil prediksi dan produksi *actual*

5. Visualisasi pentingnya *feature importance*

Selain menilai akurasi prediksi, dilakukan pula analisis *feature importance* untuk mengetahui sejauh mana kontribusi masing-masing variabel input terhadap hasil prediksi. Hasilnya

ditunjukkan pada Gambar 4, di mana variabel Luas Panen memberikan pengaruh paling besar terhadap hasil model, diikuti oleh Curah Hujan**,** Kelembapan, dan Suhu rata-rata. Hal ini menunjukkan bahwa faktor luas panen memiliki keterkaitan langsung dengan volume produksi, sedangkan variabel iklim berperan sebagai faktor pendukung yang memengaruhi produktivitas secara tidak langsung.



Gambar 4. Pentingnya masing-masing fitur dalam prediksi

3.6. *Hyperparameter tuning* dan *Cross-Validation*

Untuk meningkatkan performa model *Random Forest*, dilakukan *Hyperparameter tuning* menggunakan GridSearchCV. Parameter yang di-tuning antara lain n_estimators (jumlah pohon dalam hutan), max_depth (kedalaman maksimum pohon keputusan), dan min_samples_split (jumlah minimum sampel untuk pemisahan). Selain itu, dilakukan *Cross-Validation* (5-*fold*) untuk memastikan bahwa model tidak mengalami *overfitting* dan dapat menghasilkan generalisasi yang lebih baik. Hasil GridSearchCV menunjukkan parameter terbaik yang menghasilkan performa model optimal.

Secara rinci, tahapan yang dilakukan dapat dijelaskan sebagai berikut:

Tabel 3. Hasil *hyperparameter tuning* Manual

No	Jumlah n_*estimators*	R^2 Score
1.	50	0.8770
2.	100	0.8824
3.	200	0.8870

Berdasarkan hasil *hyperparameter* tuning yang ditampilkan pada Tabel 3, terlihat bahwa pengujian jumlah n_estimators secara manual memberikan variasi nilai R^2 score yang relatif stabil. Pada penggunaan 50 estimator, diperoleh nilai R^2 sebesar 0,8770, sementara dengan 100 estimator nilai R^2 meningkat menjadi 0,8824 yang merupakan hasil terbaik. Namun, ketika jumlah estimator ditingkatkan hingga 200, nilai R^2 justru sedikit menurun menjadi 0,8870 sehingga menunjukkan bahwa penambahan jumlah estimator yang terlalu besar tidak selalu memberikan peningkatan performa yang signifikan.

Tabel 4. Hasil *hyperparameter tuning* dengan *GridSearchCV*

No	Parameter yang Disetel	Nilai Terbaik
1.	N_*estimators*	100
2.	*Max_depth*	10
3.	*Min_samples_Split*	2

Selanjutnya, hasil *hyperparameter tun*ing menggunakan *GridSearchCV* pada Tabel 4 menunjukkan bahwa kombinasi parameter terbaik untuk model adalah jumlah n_estimators sebanyak 100, max_depth sebesar 10, dan min_samples_split bernilai 2. Hasil ini menguatkan temuan dari tuning manual bahwa 100 estimator merupakan pilihan optimal, sekaligus memberikan informasi tambahan terkait kedalaman maksimum pohon dan jumlah minimum sampel untuk pemisahan node yang lebih sesuai. Dengan demikian, *GridSearchCV* terbukti mampu memberikan hasil pencarian parameter yang lebih komprehensif dibandingkan pengujian manual.

3.7. Penyimpanan Model dan Scaler

Setelah model terbaik ditemukan, baik *Random Forest* yang telah dilatih maupun Scaler yang digunakan untuk normalisasi fitur, disimpan menggunakan joblib. Penyimpanan ini bertujuan agar model dan Scaler dapat digunakan kembali di masa mendatang untuk prediksi dengan dataset baru tanpa perlu pelatihan ulang. Model *Random Forest* yang telah dilatih dan objek Scaler disimpan menggunakan joblib. Penyimpanan ini bertujuan agar model dan proses skala data dapat langsung digunakan kembali tanpa perlu pelatihan ulang, sehingga mempercepat proses prediksi di masa depan.

3.8. Prediksi Data Baru

Untuk menguji kemampuan model dalam memprediksi produksi padi berdasarkan data baru, dilakukan simulasi Input berupa data sintetis. Data baru ini berisi informasi mengenai luas panen, curah hujan, kelembaban, dan suhu rata-rata. Data tersebut kemudian dinormalisasi menggunakan Scaler yang telah disimpan sebelumnya agar sesuai dengan skala data pelatihan. Setelah proses normalisasi, data baru diberikan ke model *Random Forest* yang telah dilatih untuk menghasilkan prediksi produksi padi dalam satuan ton.
Langkah-langkahnya adalah sebagai berikut:

1. Membuat Dataframe berisi Input baru.
2. Melakukan transformasi data menggunakan StandardScaler yang sama seperti pada saat training.
3. Menggunakan model terlatih untuk melakukan prediksi.
4. Menampilkan hasil estimasi produksi padi berdasarkan Input yang diberikan.

4. KESIMPULAN DAN SARAN

Penelitian ini dilakukan untuk menganalisis hubungan antara produksi padi dan faktor-faktor lingkungan di wilayah Sumatera, dengan mengaplikasikan pendekatan data science dan algoritma *Machine learning*. Data yang digunakan dalam analisis mencakup informasi mengenai produksi padi, luas panen, curah hujan, kelembaban, dan suhu rata-rata dari berbagai provinsi di Sumatera dalam rentang tahun 1993 hingga 2020. Temuan penelitian mengindikasikan bahwa luas panen memiliki korelasi terkuat dengan produksi padi, sementara faktor-faktor lingkungan juga memberikan pengaruh yang signifikan. Model *Random Forest* dipilih sebagai model yang paling sesuai berdasarkan evaluasi metrik R^2, MAE, dan RMSE. Penelitian ini menegaskan urgensi pemanfaatan pendekatan kuantitatif berbasis data dalam rangka mendukung perencanaan dan perumusan kebijakan pertanian yang lebih akurat.

Sejalan dengan hasil penelitian ini, direkomendasikan agar studi selanjutnya memperluas cakupan variabel penelitian dengan mempertimbangkan faktor-faktor lain yang relevan terhadap produktivitas padi, seperti jenis varietas, sistem irigasi, dan aplikasi pupuk. Lebih lanjut, pengumpulan data yang mencakup wilayah geografis dan periode waktu yang lebih ekstensif akan berkontribusi pada peningkatan generalisasi model prediksi. Pengembangan model *machine learning* juga disarankan untuk dieksplorasi lebih lanjut melalui komparasi berbagai algoritma,

serta implementasi validasi eksternal menggunakan data terkini guna meningkatkan akurasi dan aplikabilitas prediksi dalam mendukung proses pengambilan keputusan di sektor pertanian.

REFERENSI

[1] Nurzannah, S. E., Girsang, M. A., & EI Ramija, K. (2020). Faktor-faktor yang mempengaruhi produksi padi sawah (Oryza sativa L.) di Kabupaten Serdang Bedagai. Jurnal Pengkajian dan Pengembangan Teknologi Pertanian, 23(1), 11-24. https://repository.pertanian.go.id/server/api/core/bitstreams/67b0bf3d-6280-48d3-9799-cdc575b885e2/content

[2] Tarisa, D. M. H. (2022). Analisis faktor-faktor yang mempengaruhi produksi padi di Kabupaten Pati tahun 1990–2019. Jurnal Litbang Kota Pekalongan, 20(2), 107–118. https://doi.org/10.54911/litbang.v20i2.215

[3] Alfarisy, D. A., Arif, C., & Purwanto, A. (2024). Pengembangan model identifikasi air-lingkungan-tanaman untuk budidaya padi sawah dengan perlakuan fine bubble technology. Jurnal Teknik Sipil dan Lingkungan, 9(2), 231-240. https://doi.org/10.29244/jsil.9.2231-240

[4] Suryani, D. (2023). Analisis pengaruh faktor-faktor produksi terhadap produktivitas padi sawah di Kabupaten Indramayu [Skripsi, Universitas Padjadjaran]. Repositori Universitas Padjadjaran. https://repository.unpad.ac.id/server/api/core/bitstreams/0ca6169a-5a7c-43b0-884c-213d12fbdc5a/content

[5] Mila. (2019). Analisis usaha Kipang Andalas di Kota Payakumbuh [Skripsi, Universitas Andalas]. Repositori Universitas Andalas. https://scholar.unand.ac.id/44449/

[6] Auliya, D., Rosandi, A. H., & Subroto, W. T. (2024). Analisis perubahan iklim terhadap produktivitas padi di Jawa Timur. Diponegoro Journal of Economics, 13(3), 55–65. https://doi.org/10.14710/djoe.47595

[7] Breiman,L.(2001). *Random Forest*s. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[8] Wahid, A., Rauf, A., & Syarifuddin, S. (2023). Analisis prediksi produksi padi menggunakan metode *Machine learning* di Indonesia. Jurnal Ilmiah Teknologi Informasi Terapan, 9(2), 75–83. https://doi.org/10.35308/jitit.v9i2.5467

[9] Setiyono, T. D., Ali, A. M., Quilang, J. P., & Barredo, J. I. (2019). Rice yield estimation using satellite remote sensing and *Machine learning* models. Remote Sensing, 11(7), 752. https://doi.org/10.3390/rs11070752

[10] Food and Agriculture Organization. (2017). The future of food and agriculture: Trends and challenges. FAO. https://www.fao.org/3/i6583e/i6583e.pdf