Journal of Computer Science and Computational Intelligence
Mei 2025 | Vol. 1 | No. 1 | Page 0009-0019

Prediksi Produktivitas Padi Menggunakan Algoritma *Random Forest* di Provinsi Sumatera Tahun 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**

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 Sumatra 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, MAE, and RMSE 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.*

1. PENDAHULUAN

1.1. Latar Belakang

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 [3]. 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 [4]. 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 [2]. 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. Ketiga faktor lingkungan ini tidak dapat dikendalikan secara langsung oleh petani, namun sangat mempengaruhi hasil panen. Variabilitas iklim yang semakin tinggi akibat perubahan iklim global juga menambah ketidakpastian dalam produksi pertanian [6]. Di sisi lain, perkembangan teknologi data science memberikan peluang baru dalam memahami dinamika produksi padi secara lebih kuantitatif dan objektif. Dengan memanfaatkan data historis yang mencakup berbagai variabel produksi dan iklim, kita dapat melakukan analisis statistik dan membangun model prediktif untuk memperkirakan hasil produksi padi di masa mendatang [3]. Pemanfaatan pendekatan ini tidak hanya akan memperkaya wawasan dalam bidang pertanian, tetapi juga menjadi landasan penting dalam penyusunan kebijakan pertanian yang berbasis data (data-driven policy). Terlebih lagi, dengan kemampuan untuk memprediksi hasil produksi, para pemangku kebijakan dapat mengambil langkah antisipatif dalam memastikan stabilitas pangan dan kesejahteraan petani [2] [4]. 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.

1.2. Tinjauan Pustaka

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. Faktor lingkungan seperti curah hujan, suhu, dan kelembaban udara memberikan pengaruh signifikan terhadap hasil produksi pertanian, termasuk padi [1]. Sejalan dengan itu, perkembangan teknologi analitik seperti *machine learning* memberikan pendekatan baru dalam pemodelan prediksi hasil panen. Pendekatan sistem air-lingkungan-tanaman dikembangkan untuk memahami keterkaitan antara variabel lingkungan dan hasil pertanian [3]. 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 [11]. 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. METODOLOGI

2.1. Pemrosesan Data

Dataset yang digunakan dalam penelitian ini terdiri dari 224 entri data yang mencakup beberapa provinsi di Sumatera dari tahun 1993 hingga 2020. Fitur-fitur yang tersedia dalam dataset ini meliputi:

1. Provinsi: Nama provinsi yang mencakup wilayah penelitian.
2. Tahun: Tahun pengamatan data.
3. Produksi (ton): Produksi padi yang dihasilkan dalam satuan ton.

4. Luas Panen (hektar): Luas lahan yang digunakan untuk penanaman padi.
5. curah hujan (mm): Total curah hujan yang terjadi selama periode tersebut.
6. Kelembaban (%): Tingkat kelembaban rata-rata selama periode tersebut.
7. Suhu Rata-rata (°C): Suhu rata-rata selama periode tersebut.

Langkah-langkah pemrosesan data mencakup:

1. Pemeriksaan Nilai Kosong: Memastikan bahwa dataset tidak memiliki missing values.
2. Standarisasi Nama Kolom: Untuk mempermudah analisis dan konsistensi.
3. Visualisasi Distribusi Data dan Deteksi *Outlier*: Menggunakan Boxplot untuk mendeteksi adanya *Outlier* dalam data yang dapat mempengaruhi kualitas model.
4. Normalisasi Data: Fitur numerik distandarisasi menggunakan StandardScaler untuk memastikan setiap fitur memiliki skala yang sama, menghindari dominasi 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:

1. Rata-rata curah hujan per Tahun per Provinsi: Merupakan agregasi tahunan untuk melihat pola curah hujan sepanjang tahun.
2. Perubahan Suhu dari Tahun ke Tahun: Menganalisis fluktuasi suhu sebagai faktor yang dapat mempengaruhi pertumbuhan padi.
3. Faktor Indeks Lingkungan: Gabungan dari 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).

Dataset kemudian dibagi menjadi dua bagian: fitur (X) dan target (y), dengan target berupa produksi padi.

2.4. Pembuatan Model Prediksi

Model prediksi dibuat menggunakan algoritma regresi, dengan pilihan model sebagai berikut:

1. Linear regression: Sebagai model dasar untuk analisis hubungan linier.
2. *Random Forest regressor*: Sebagai model *ensemble* yang lebih kuat untuk menangani data dengan variabilitas tinggi.

3. Decision Tree Regressor: Sebagai model non-linear yang mudah diinterpretasi.

Proses pelatihan dan evaluasi model dilakukan dengan langkah-langkah berikut:

1. Pembagian Data: Dataset dibagi menjadi data latih (80%) dan data uji (20%).
2. Pelatihan Model: Model dilatih pada data latih untuk memahami hubungan antara fitur dan target.
3. Evaluasi Model: Menggunakan metrik R^2, MAE, dan RMSE untuk menilai seberapa baik model dapat memprediksi produksi padi pada data uji.
4. Pemilihan Model Terbaik: Berdasarkan skor evaluasi yang diperoleh dari model-model yang diuji.

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

1. Model prediksi dibangun menggunakan beberapa algoritma regresi:
2. *Linear regression*: Sebagai model dasar untuk analisis hubungan linier.
3. *Random Forest regressor*: Sebagai model *ensemble* yang lebih kuat untuk menangani data dengan variabilitas tinggi.
4. *Decision Tree Regressor*: Sebagai model non-linear yang mudah diinterpretasi.

Langkah-langkah yang dilakukan adalah:

1. Split data: Data dibagi menjadi data latih (80%) dan data uji (20%) menggunakan metode train_test_Split.
2. Training:Model dilatih pada data latih untuk memahami hubungan antara fitur dan target.
3. Evaluasi: Model dievaluasi menggunakan metrik R^2, MAE (*Mean Absolute Error*), dan RMSE (*Root Mean Squared Error*) untuk menilai seberapa baik model dapat memprediksi produksi padi pada data uji.
4. Pemilihan Model Terbaik: Model yang memberikan skor evaluasi tertinggi dipilih dari sejumlah model yang diuji dan akan digunakan untuk 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:

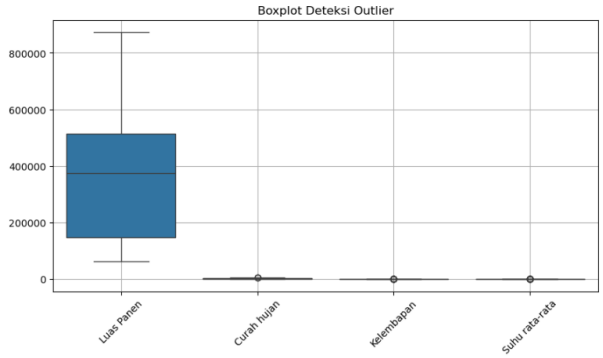
1. Load Dataset. Pada bagian ini, berbagai library yang diperlukan untuk analisis dipanggil, seperti pandas untuk manipulasi data, numpy untuk operasi numerik, matplotlib dan seaborn untuk visualisasi, serta library dari sklearn untuk pemodelan dan evaluasi. Dataset kemudian dimuat ke dalam sebuah DataFrame 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 di-run 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.

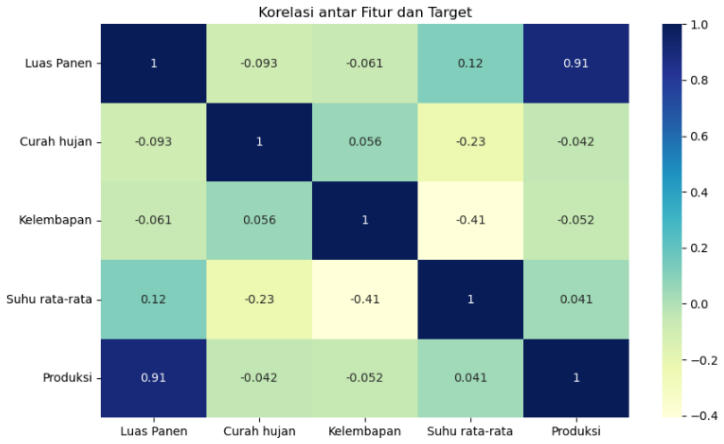


Gambar 1: Deteksi *outlier* pada fitur numerik menggunakan boxplot.

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

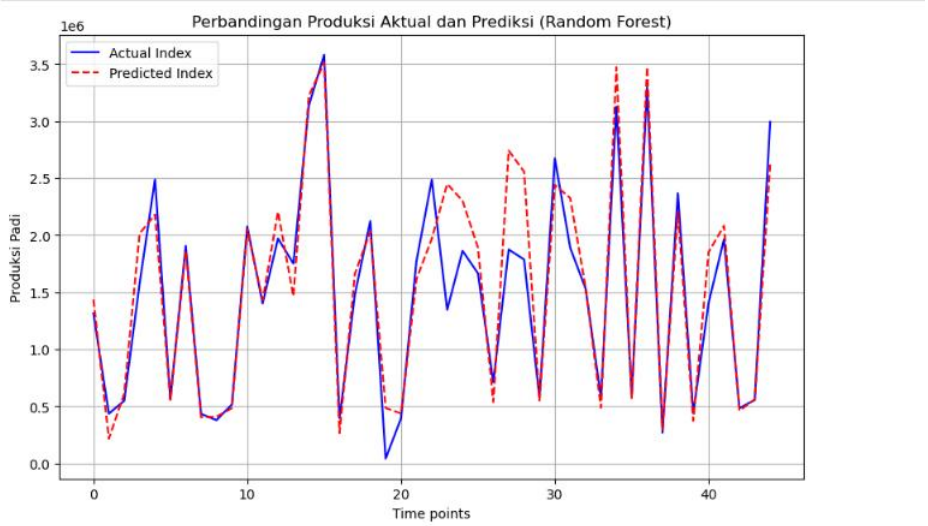
Tabel 1. Hasil Evaluasi Model pada Data Uji

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

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

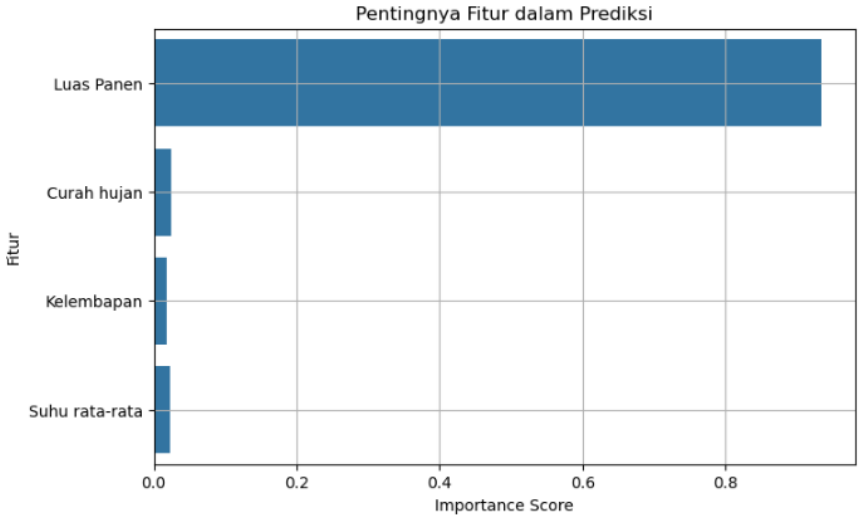
No	Metode Evaluasi	Nilai
1.	*R^2 Score Cross-Validation (5-Fold)*	-1.4581 ± 4.0399

4. Visualisasi hasil prediksi (dengan grafik garis)



Gambar 3. Perbandingan hasil prediksi dan produksi actual.

5. Visualisasi pentingnya fitur Importances



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

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

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:

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

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