Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

Prediksi Produksi Tanaman Padi di Indonesia dengan Menggunakan Algoritma Random Forest Regressor

Dionikxon Manurung[1], Billiam Zealtiel[2], Andre Hasudungan Lubis[3,*]

Fakultas Teknik, Teknik Informatika, Universitas Medan Area, Kota Medan, Indonesia
Email: [1]dionikxonmanurung@gmail.com, [2]billiamzealtiel1@gmail.com, [3,*]andrelubis2201@gmail.com
Email Penulis Korespondensi: andrelubis2201@gmail.com

Abstrak−Produksi padi merupakan komponen utama dalam menjaga ketahanan pangan nasional di Indonesia, mengingat beras adalah makanan pokok mayoritas penduduk. Namun, kestabilan produksi padi sering kali terganggu oleh berbagai faktor, terutama kondisi agronomis dan variabilitas iklim yang sulit diprediksi. Oleh karena itu, diperlukan pendekatan berbasis data yang mampu memodelkan kompleksitas faktor-faktor tersebut secara akurat. Penelitian ini bertujuan untuk membangun model prediksi produksi padi menggunakan algoritma *Random Forest Regressor*, sebuah metode pembelajaran mesin yang dikenal andal dalam menangani data non-linear dan kompleks. Dataset yang digunakan mencakup parameter pertanian seperti luas panen dan produktivitas, serta data iklim meliputi suhu, kelembaban udara, dan curah hujan, yang dikumpulkan dari sumber terbuka seperti Kaggle dan Badan Meteorologi Klimatologi dan Geofisika (BMKG) untuk rentang tahun 2018 hingga 2024. Metodologi yang diterapkan dalam penelitian ini terdiri dari beberapa tahapan, yaitu prapemrosesan data (penanganan nilai hilang dan normalisasi), analisis data eksploratif untuk memahami pola dan korelasi antar variabel, pelatihan model prediksi, serta evaluasi performa model menggunakan metrik *Mean Squared Error (MSE)* dan *R-squared (R^2)*. Hasil penelitian menunjukkan bahwa konfigurasi terbaik diperoleh saat data dibagi dengan rasio pelatihan dan pengujian sebesar 90:10, serta penggunaan 200 *decision tree* dalam model. Konfigurasi ini menghasilkan nilai MSE sebesar 0.0004 dan R^2 sebesar 0.9918, yang mengindikasikan tingkat akurasi prediksi yang sangat tinggi serta kemampuan model dalam merepresentasikan hubungan antar variabel dengan baik. Penelitian ini menunjukkan bahwa *Random Forest Regressor* efektif dalam memprediksi produksi padi dan berpotensi menjadi alat bantu pengambilan keputusan strategis bagi pemangku kepentingan di sektor pertanian.

Kata Kunci : Prediksi Padi; *Machine Learning*; *Random Forest Regressor*

Abstract−*Rice production is a key component in maintaining national food security in Indonesia, given that rice is the staple food for the majority of the population. However, the stability of rice production is often disrupted by various factors, particularly agronomic conditions and unpredictable climate variability. Therefore, a data-driven approach capable of accurately modeling the complexity of these factors is needed. This study aims to build a rice production prediction model using the Random Forest Regressor algorithm, a machine learning method known for its reliability in handling non-linear and complex data. The dataset used includes agricultural parameters such as harvested area and productivity, as well as climate data including temperature, air humidity, and rainfall, collected from open sources such as Kaggle and the Meteorology, Climatology, and Geophysics Agency (BMKG) for the period 2018 to 2024. The methodology applied in this study consists of several stages, namely data preprocessing (handling missing values and normalization), exploratory data analysis to understand patterns and correlations between variables, training a prediction model, and evaluating model performance using the Mean Squared Error (MSE) and R-squared (R^2) metrics. The results show that the best configuration is obtained when the data is divided with a training and testing ratio of 90:10, and the use of 200 decision trees in the model. This configuration produces an MSE value of 0.0004 and an R^2 of 0.9918, which indicates a very high level of prediction accuracy and the model's ability to represent the relationship between variables well. This study demonstrates that the Random Forest Regressor is effective in predicting rice production and has the potential to be a strategic decision-making tool for stakeholders in the agricultural sector.*

Keywords : *Rice Prediction; Machine Learning; Random Forest Regressor; MSE; R-squared*

1. PENDAHULUAN

Sektor pertanian merupakan salah satu pilar penting penggerak perekonomian Indonesia, sejalan dengan statusnya sebagai negara agraris. Di antara berbagai komoditas pertanian, padi menempati posisi yang sangat strategis karena menjadi sumber pangan utama bagi lebih dari 270 juta penduduk Indonesia [1]. Ketersediaan dan stabilitas produksi padi menjadi penting tidak hanya berdampak pada kesejahteraan petani, tetapi juga berperan penting dalam menjaga ketahanan pangan nasional dan mengendalikan dinamika pasar, baik domestik maupun internasional. Oleh karena itu, upaya peningkatan produktivitas serta efisiensi distribusi hasil panen padi terus menjadi perhatian utama dalam kebijakan pertanian nasional.

Namun demikian, produksi padi masih menghadapi ketakstabilan yang dipengaruhi oleh beragam faktor, seperti, perubahan luas panen, tingkat produktivitas, suhu, dan curah hujan. Hal tersebut membuat perencanaan distribusi hasil panen dan kebijakan stok beras nasional menjadi kurang optimal jika hanya mengandalkan data historis tanpa pendekatan prediktif. Oleh karena itu, prediksi produksi padi yang menggunakan variabel pertanian dan iklim menjadi penting untuk memastikan kesiapan sektor pertanian menghadapi dinamika tersebut. Dalam konteks ini, kemampuan untuk melakukan prediksi berbasis data historis beragam variabel menjadi sangat bernilai. Kemampuan untuk menyediakan informasi prediktif memungkinkan pihak-pihak terkait, seperti petani, distributor, dan pemerintah, untuk menyusun rencana kerja dan kebijakan yang lebih tepat sasaran. Prediksi yang akurat dapat membantu memperkirakan ketersediaan hasil panen pada periode tertentu, menjaga kesembangan antara produksi, serta menyesuaikan kebutuhan logistik dan pemasaran. Dengan fokus pada variabel yang relevan dan mudah diakses, model prediksi menjadi lebih sederhana namun tetap memberikan informasi yang yang sangat diperlukan untuk mendukung pengambilan keputusan di sektor pertanian [2].
Penelitian tentang prediksi produksi padi telah banyak dilakukan dengan menggunakan berbagai algoritma, termasuk regresi linier, ARIMA, dan model machine learning seperti Random Forest dan LSTM. Salah satu studi menunjukkan

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

bahwa regresi linier dapat menjelaskan sekitar 82,6% variasi produksi padi dengan menggunakan variabel seperti luas panen, curah hujan, kelembaban, dan suhu rata-rata. Hal ini sejalan dengan penelitian sebelumnya yang juga menggunakan regresi linier untuk memprediksi produksi tanaman padi di Kabupaten Grobogan, Jawa Tengah, yang menghasilkan akurasi prediksi yang cukup baik. Namun, model regresi linier memiliki keterbatasan dalam menangani pola musiman dan fluktuasi cuaca, yang sering menjadi tantangan utama dalam pertanian. Penggunaan ARIMA menunjukkan keunggulan dalam menangkap pola musiman pada data pertanian, meskipun model ini tidak seakurat model berbasis deep learning seperti LSTM. Penelitian lain menguji perbandingan antara MLPRegressor dan LSTM untuk memprediksi produktivitas padi di Indonesia, dan hasilnya menunjukkan bahwa LSTM dengan arsitektur 2-42-42-42-1 memiliki akurasi prediksi lebih tinggi (94,12%) dibandingkan MLPRegressor yang hanya mencapai 91,18%, mengindikasikan bahwa LSTM lebih unggul dalam menangkap pola musiman yang kompleks dalam data pertanian. Selain itu, analisis terhadap prediksi gagal panen menggunakan berbagai algoritma machine learning seperti Random Forest dan XGBoost menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi mencapai 77%, lebih baik daripada metode lainnya [3].

Salah satu metode alternatif yang menunjukkan kinerja unggul dalam berbagai kasus prediksi adalah Random Forest Regressor. Algoritma ini merupakan bagian dari pendekatan *ensemble learning* yang menggabungkan sejumlah decision tree (decision trees) untuk menghasilkan prediksi yang lebih stabil dan akurat. Keunggulan utama dari Random Forest terletak pada kemampuannya dalam menangani data dengan dimensi tinggi, hubungan non-linear antar variabel, serta keberadaan noise atau *outlier* dalam dataset. Selain itu, metode ini tidak memerlukan asumsi distribusi data seperti pada pendekatan statistik konvensional, sehingga lebih fleksibel dalam mengolah data yang bervariasi seperti data produksi padi. Dengan sistem pemungutan suara dari banyak pohon, Random Forest juga lebih tahan terhadap overfitting dibandingkan dengan model prediktif tunggal. Karakteristik ini menjadikan Random Forest sebagai kandidat yang menjanjikan untuk diterapkan dalam prediksi hasil produksi padi yang kompleks dan dinamis [4],[5].

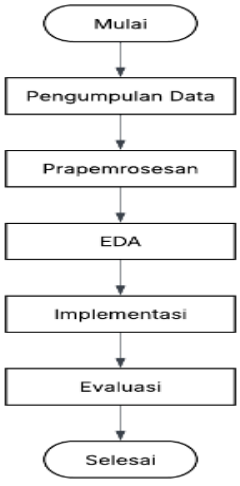
Saat ini Random Forest telah terbukti memberikan performa yang baik dalam berbagai domain seperti prediksi yang berkaitan dengan penyakit dalam dunia kesehatan, pasar finansial, hingga klasifikasi berita hoax. Namun penerapannya dalam konteks prediksi produksi padi di Indonesia masih jarang ditemukan. Padahal, struktur data pertanian yang multivariat dan dipengaruhi oleh faktor lingkungan yang tidak sepenuhnya terkontrol sangat sesuai untuk ditangani oleh pendekatan Random Forest. Selain itu, model ini juga menawarkan keunggulan dalam hal interpretabilitas relatif, di mana pentingnya setiap variabel prediktor dapat diukur dan dianalisis. Dengan mempertimbangkan kompleksitas variabel yang mempengaruhi hasil panen padi—seperti curah hujan, suhu, kelembapan, luas tanam, dan jenis benih—Random Forest berpotensi menjadi solusi prediktif yang akurat dan aplikatif di lapangan [6].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun model prediksi hasil produksi padi menggunakan algoritma Random Forest Regressor. Model yang dikembangkan akan dilatih dan diuji menggunakan data historis produksi padi yang mencakup berbagai parameter agronomis dan lingkungan. Evaluasi kinerja model dilakukan dengan menggunakan metrik statistik seperti Mean Square Error (MSE) dan koefisien determinasi (R^2). Dengan pendekatan ini, diharapkan model yang dihasilkan tidak hanya mampu memberikan prediksi yang akurat, tetapi juga dapat menjadi alat bantu pengambilan keputusan yang efektif bagi pemangku kepentingan di sektor pertanian [7].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama yang sistematis untuk mencapai tujuan prediksi produksi tanaman padi di Indonesia dengan menggunakan algoritma Random Forest Regressor. Tahapan-tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1. Tahapan Penelitian

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

a. Tahapan pertama dilakukan pengumpulan data
   Pengumpulan data dilakukan dengan memanfaatkan dataset sekunder yang diperoleh dari situs web Kaggle. Dataset yang terkumpul dapat dikategorikan kedalam dua bagian utama, yakni data pertanian dan data iklim.
b. Prapemrosesan Data
   Selanjutnya, data yang telah dikumpulkan masuk ke dalam tahap prapemrosesan. Pada tahap ini, dilakukan pemeriksaan terhadap data untuk mendeteksi adanya nilai yang hilang (*missing value*) yang berpotensi mempengaruhi kualitas analisis. Selain itu, dilakukan proses *feature engineering* untuk mengolah dan menciptakan fitur baru yang dapat meningkatkan performa model prediksi. Data juga dinormalisasi menggunakan metode Min-Max Scaler, yakni metode yang menggunakan nilai minimum dan nilai maksimum untuk menyetarakan skala antar data sehingga berada pada rentang yang sama. Normalisasi ini mengikuti rumus:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

c. Exploratory Data Analysis (EDA)
   Exploratory Data Analysis (EDA) adalah proses awal dalam analisis data yang bertujuan untuk memahami struktur, karakteristik, dan pola dalam dataset sebelum dilakukan pemodelan. Pada penelitian ini, EDA dilakukan untuk menggali informasi penting dari dataset produksi padi di Indonesia sebelum diterapkannya algoritma prediksi.
d. Implementasi
   mplementasi penelitian ini menggunakan algoritma Random Forest

2.2 Algoritma Random Forest

Algoritma Random Forest Regressor merupakan salah satu metode ensemble learning yang bekerja dengan membangun banyak decision tree (*decision trees*) dan menggabungkan hasil prediksi dari masing-masing pohon untuk menghasilkan output akhir. Algoritma ini termasuk dalam kategori *supervised learning* yang digunakan untuk menyelesaikan permasalahan regresi, di mana nilai keluaran bersifat kontinu. Keunggulan dari *Random Forest* terletak pada kemampuannya dalam mengurangi *overfitting*, meningkatkan akurasi, dan bekerja dengan baik terhadap data berdimensi tinggi [8],[9],[10]. Pada prediksi, algoritma *Random Forest Regressor* berfungsi untuk mempelajari pola dan hubungan dari berbagai fitur (variabel independen) terhadap nilai target yang ingin diprediksi (dalam hal ini, produksi padi). Sebelum model diterapkan, penting untuk memahami keterkaitan antar variabel dalam dataset. Variabel produksi berperan sebagai target atau variabel dependen, sementara luas panen dan produktivitas merupakan fitur atau variabel independen. Produksi padi secara logis dan matematis merupakan hasil dari interaksi antara luas panen dan produktivitas, sehingga hubungan di antara ketiga variabel ini sangat erat dan bersifat deterministik [6]. Korelasi antara fitur dan target tersebut menjadi acuan dalam proses pelatihan model, karena semakin kuat korelasi, semakin tinggi kemungkinan fitur tersebut memiliki pengaruh signifikan. Setelah hubungan antar variabel diketahui, proses pelatihan model dilakukan dengan memasukkan fitur terpilih ke dalam algoritma regresi [11],[12].

$$\hat{Y}i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{Y}\, n \quad (2)$$

Dalam implementasi pada kasus produksi padi, algoritma *Random Forest Regressor* dilatih menggunakan data historis produksi serta berbagai fitur yang memengaruhi hasil panen. Pada tahap ini juga akan dilakukan fine-tuning yakni proses mengadaptasi parameter model untuk meningkatkan performanya. Setelah model dilatih, ia dapat digunakan untuk memprediksi produksi padi di masa mendatang berdasarkan nilai fitur yang tersedia, seperti kondisi iklim, luas tanam, dan jenis varietas. Model ini dapat memberikan informasi yang berguna bagi pemerintah, dan petani untuk merencanakan strategi pertanian yang lebih efektif dan adaptif. Selain itu, dengan mengetahui fitur mana yang paling berpengaruh terhadap produksi, intervensi dapat dilakukan secara lebih tepat sasaran. Pada tahap ini juga akan dilakukan fine-tuning yakni proses mengadaptasi parameter model untuk meningkatkan performanya.

Mean Squared Error (MSE) adalah metrik yang digunakan dalam mengukur rata-rata dari kuadrat selisih antara nilai aktual dan nilai prediksi untuk memberikan gambaran tentang seberapa besar kesalahan yang dihasilkan oleh model secara umum (Lai et al., 2023). Semakin kecil nilai MSE, semakin dekat prediksi model dengan nilai aktual. Karena MSE menghitung kuadrat dari selisih, maka kesalahan besar akan memberi kontribusi lebih besar dalam perhitungan total kesalahan, sehingga model lebih sensitif terhadap outlier. Hal ini membuat MSE cocok digunakan dalam kasus yang membutuhkan presisi tinggi dalam prediksi.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

Proses perhitungan melibatkan selisih antara setiap nilai aktual dan prediksinya, kemudian dikuadratkan agar semua nilai positif, lalu dirata-ratakan. Nilai MSE yang besar menunjukkan bahwa prediksi model cenderung menyimpang jauh dari nilai yang sebenarnya. Sebaliknya, nilai yang kecil menunjukkan bahwa model memiliki akurasi prediksi yang tinggi secara keseluruhan. Dalam konteks prediksi produksi padi, MSE digunakan untuk mengevaluasi sejauh mana model mampu mengestimasi nilai produksi berdasarkan dua variabel masukan, yaitu luas panen dan produktivitas. Karena hubungan antara ketiga variabel tersebut bersifat hampir deterministik, model yang efektif seharusnya menghasilkan nilai

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

MSE yang rendah. Nilai MSE yang rendah menunjukkan bahwa hasil prediksi tidak menyimpang jauh dari data aktual produksi. Oleh karena itu, MSE menjadi indikator utama untuk mengukur keakuratan kuantitatif dari model ini.

R-squared (R^2), atau dikenal sebagai koefisien determinasi, merupakan metrik evaluasi regresi yang menunjukkan seberapa besar proporsi variasi pada variabel target (produksi) yang dapat dijelaskan oleh variabel input (luas panen, produktivitas, suhu, kelembaban, dan curah hujan). Evaluasi ini mempertimbangkan distribusi nilai sebenarnya (*ground truth*) dan memberikan skor tinggi apabila sebagian besar nilai hasil prediksi mendekati nilai aktual (Chicco et al., 2021). Nilai R^2 berada dalam rentang 0 hingga 1, dengan nilai mendekati 1 menunjukkan bahwa model sangat baik dalam menjelaskan variabilitas data. Metrik ini sangat berguna untuk menilai kualitas model secara proporsional.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

Dalam penelitian ini, nilai R^2 digunakan untuk menilai seberapa baik model Random Forest Regressor dapat menjelaskan hubungan antar variabel, yaitu antara luas panen dan produktivitas terhadap produksi padi. Karena secara logika produksi merupakan hasil dari interaksi dua variabel tersebut, nilai R^2 yang tinggi akan mengindikasikan bahwa model telah berhasil memetakan hubungan tersebut secara akurat. Evaluasi ini tidak hanya menunjukkan akurasi model, tetapi juga memperkuat bukti bahwa data dan fitur yang digunakan memiliki keterkaitan yang kuat [13], [14].

2.3 Marchine Learning

Machine Learning adalah cabang dari kecerdasan buatan (Artificial Intelligence) yang berfokus pada pengembangan algoritma dan model yang memungkinkan sistem komputer untuk belajar secara otomatis dari data tanpa harus diprogram secara eksplisit. Melalui proses pelatihan (training), sistem mempelajari pola dan hubungan dalam data untuk kemudian digunakan dalam membuat prediksi, pengambilan keputusan, atau pengenalan pola pada data baru secara mandiri [15], [16],[17],[18].

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Pada Tabel 1 berikut merupakan Data yang dikumpulkan merupakan data yang diambil dalam kurun waktu pertahun, mulai tahun 2018 sampai tahun 2024.

Tabel 1. Data Pertanian

Provinsi	Luas Panen (ha) 2018	...	Luas Panen (ha) 2024	Produktivitas (ku/ha) 2018	...	Produktivitas (ku/ha) 2024	Produksi (ton) 2018	...	Produksi (ton) 2024
Aceh	329515.78	...	301196.35	56.49	...	55.11	1861567,1	...	1659966,3
Sumatera Utara	408176.45	...	419463.48	51.65	...	52.56	2108284,7	...	2204875,5
Sumatera Barat	313050.82	...	295278.98	47.37	...	45.94	1483076,5	...	1356467,9
Riau	71448.08	...	56421.96	37.28	...	39.36	266375,5	...	222055,7
Jambi	86202.68	...	61625.68	44.44	...	45.6	383045,7	...	281022,1
...	...	...	...	...	...	...	...	...	
P. Barat Daya	-	...	363.87	-	...	27.17	-	...	988,6
Papua	52411.95	...	1068.57	42.57	...	43.14	223119,4	...	4610,0
Papua Selatan	-	...	47168.57	-	...	46.17	-	...	217789,6
Papua Tengah	-	...	1436.12	-	...	42.28	-	...	6072,4

Selanjutnya dapat dilihat pada tabel 2 yaitu Data Iklim.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

Tabel 2. Data Iklim

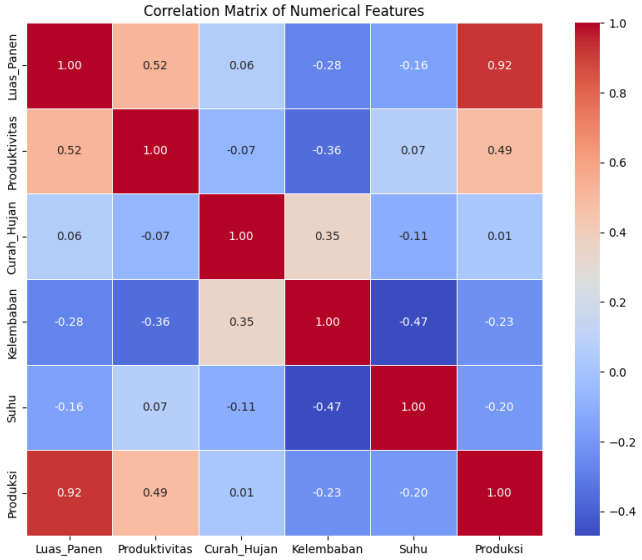
Provinsi	Curah Hujan 2018 (mm/tahun)	...	Curah Hujan 2024 (mm/tahun)	Suhu 2018 (°C)	...	Suhu 2024 (°C)	Kelembaban 2018 (%)	...	Kelembaban 2024 (%)
Aceh	2336	...	1505	28	...	28	81	...	84
Sumatera Utara	2388	...	3686	29	...	29	79	...	80
Sumatera Barat	3823	...	6507	27	...	27	87	...	85
Riau	2699	...	3101	27	...	28	81	...	81
Jambi	2643	...	3173	27	...	28	85	...	86
...	...	...	...	...	...	...	...	...	
P. Barat Daya	-	...	-	-	...	-	-	...	-
Papua	2991	...	3240	28	...	28	83	...	85
Papua Selatan	-	...	-	-	...	-	-	...	-
Papua Tengah	-	...	-	-	...	-	-	...	-
P. Pegunungan	-	...	-	-	...	-	-	...	-

3.2 Prapemrosesan

Prapemrosesan data merupakan langkah krusial sebelum model machine learning dapat dilatih secara optimal. Pada penelitian ini, proses prapemrosesan dilakukan melalui tiga tahapan utama. Tahap pertama adalah penanganan *missing value*. Pemeriksaan awal terhadap dataset menunjukkan adanya nilai kosong pada beberapa variabel numerik. Untuk menjaga integritas data tanpa mengurangi jumlah sampel yang tersedia, nilai-nilai yang hilang tersebut diatasi menggunakan metode imputasi rata-rata (mean imputation). Pendekatan ini dipilih karena bersifat sederhana namun efektif dalam mempertahankan distribusi data numerik yang tidak terlalu ekstrem. Tahap kedua adalah penghapusan kolom-kolom yang tidak berkontribusi langsung terhadap proses prediksi, yakni kolom *provinsi* dan *tahun*. Kedua kolom ini bertindak sebagai *identifier* dan tidak merepresentasikan informasi numerik yang dapat digunakan sebagai fitur prediktor dalam model regresi. Dengan menghapus kolom tersebut, dimensi dataset dapat direduksi dan kompleksitas model pun dapat ditekan tanpa mengorbankan informasi penting. Tahap ketiga adalah normalisasi data menggunakan metode *Min-Max Scaling*. Normalisasi diperlukan karena rentang nilai antar fitur berbeda-beda, misalnya luas panen yang dapat mencapai jutaan satuan, sedangkan kelembaban dan suhu berada dalam skala puluhan. Teknik Min-Max Scaling mengubah semua fitur numerik ke dalam rentang 0 hingga 1, sehingga mencegah dominasi fitur dengan skala besar terhadap proses pelatihan model. Normalisasi ini juga memastikan bahwa algoritma Random Forest dapat bekerja secara seimbang pada setiap fitur tanpa bias akibat perbedaan skala.

3.2 EDA

Tahapan *Exploratory Data Analysis (EDA)* dilakukan untuk memahami struktur dan hubungan antarvariabel dalam dataset sebelum proses pelatihan model dilakukan. Analisis ini bertujuan mengidentifikasi pola, korelasi, serta distribusi nilai yang mungkin memengaruhi hasil prediksi model. Gambar 2 merupakan visualisasi dari analisis correlation matrix untuk variabel luas panen, produktivitas, curah hujan, kelembaban, suhu, dan produksi.

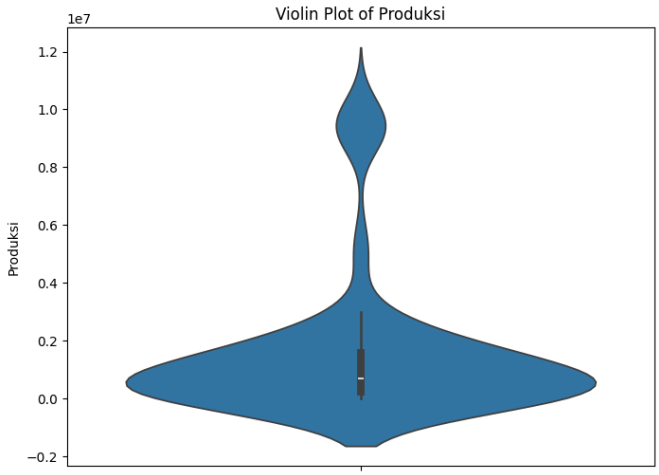


Gambar 2 Correlation Matrix

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

Dari hasil analisis korelasi tersebut, ditemukan bahwa variabel Luas Panen memiliki korelasi paling kuat terhadap Produksi, dengan nilai koefisien korelasi sebesar 0.92. Ini menunjukkan bahwa semakin luas area panen, semakin besar pula hasil produksi padi yang diperoleh, sehingga menjadikan variabel ini sebagai salah satu prediktor paling penting. Sementara itu, variabel Produktivitas juga menunjukkan korelasi positif terhadap produksi (0.49), meskipun kekuatannya lebih rendah. Di sisi lain, variabel-variabel iklim seperti Curah Hujan, Kelembaban, dan Suhu menunjukkan korelasi yang cenderung lemah terhadap produksi, dengan nilai korelasi masing-masing sebesar 0.01, -0.23, dan -0.20. Hal ini mengindikasikan bahwa meskipun faktor iklim penting dalam konteks pertanian, dalam dataset ini pengaruh langsungnya terhadap produksi padi tidak terlalu dominan secara linear.

Gambar 3 merupakan violin plot plot nilai produksi yang menggambarkan distribusi nilai, baik secara keseluruhan atau pada rentang waktu tertentu.

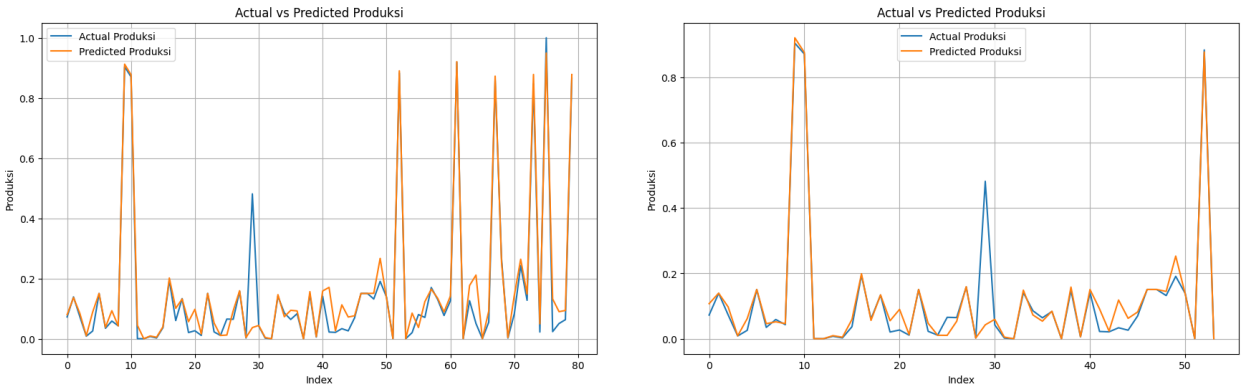


Gambar 3 Violin Plot

Violin plot pada Gambar 3 ini memberikan gambaran visual mengenai sebaran dan kepadatan data produksi. Hasil visualisasi menunjukkan bahwa sebagian besar nilai produksi terkonsentrasi pada kisaran rendah, dengan bentuk violin yang melebar di bagian bawah. Namun terdapat pula sejumlah nilai produksi yang sangat tinggi (di atas 8 juta satuan), yang memunculkan ekor panjang pada bagian atas grafik. Distribusi ini bersifat right-skewed, yang menandakan adanya outlier atau nilai-nilai ekstrem pada data produksi

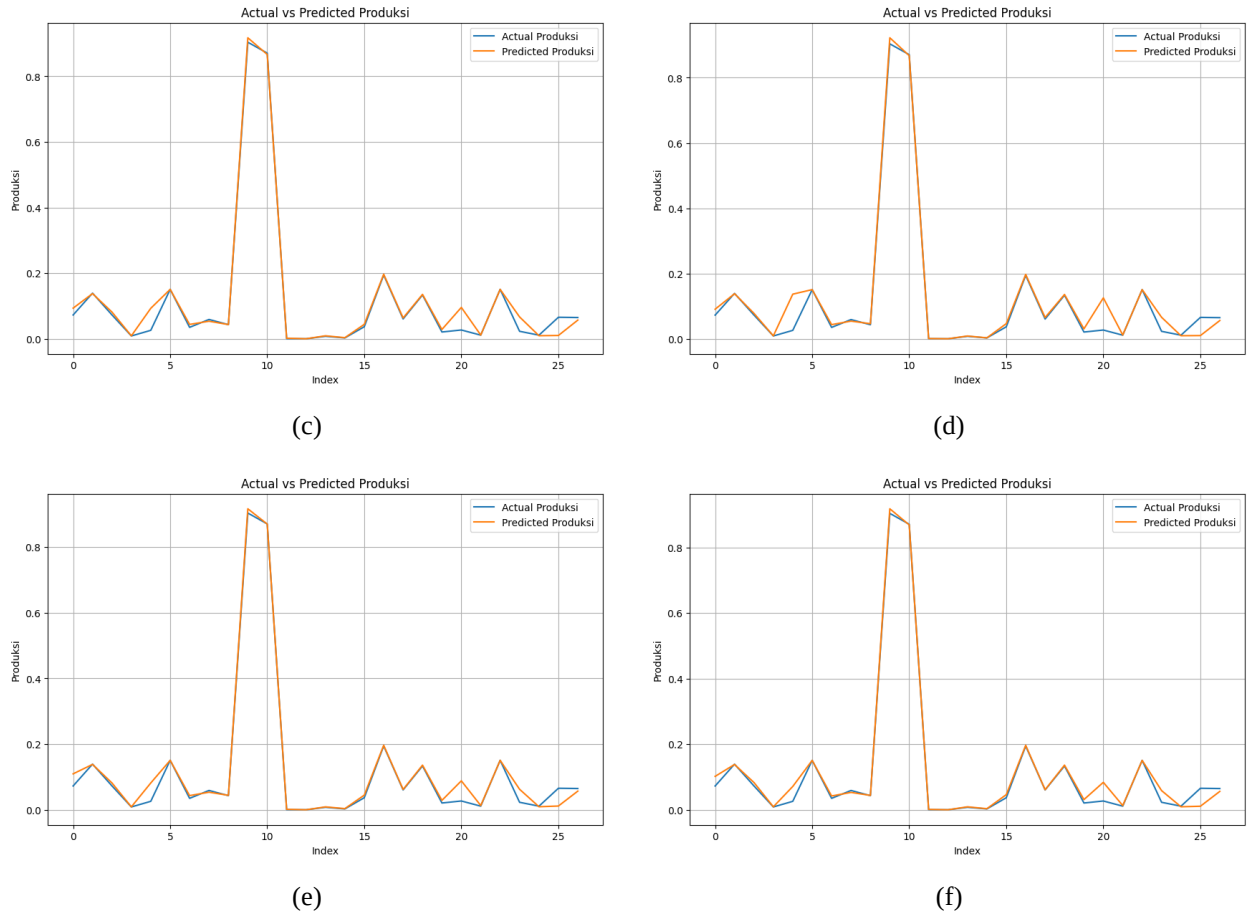
3.3 Implementasi

Setelah melalui tahapan prapemrosesan dan eksplorasi data, model *Random Forest Regressor* diimplementasikan untuk memprediksi hasil produksi padi berdasarkan variabel-variabel input seperti luas panen, produktivitas, curah hujan, suhu, dan kelembaban. Proses implementasi dilakukan menggunakan pendekatan evaluasi melalui metrik *Mean Squared Error (MSE)* dan *R-squared (R^2)* untuk menilai akurasi model. Tiga skenario pembagian data diuji, yaitu 70:30, 80:20, dan 90:10 dengan jumlah decision tree 100. Selain itu, dilakukan tuning terhadap jumlah decision tree dalam model Random Forest, dengan variasi nilai 50, 100, 150, dan 200. Setiap kombinasi parameter kemudian diuji untuk melihat pengaruhnya terhadap kualitas prediksi. Gambar 4 merupakan liine plot untuk masing masing kondisi pelatihan.



(a) (b)

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125



Gambar 4 Visualisasi hasil prediksi produksi padi oleh model Random Forest Regressor.

Pada visualisasi dengan pembagian data 70:30 (Gambar a), tampak adanya fluktuasi tajam antara kurva prediksi dan kurva aktual, khususnya pada titik-titik dengan nilai produksi tinggi. Hal ini mengindikasikan bahwa model mengalami kesulitan dalam memprediksi nilai ekstrem, yang kemungkinan besar merupakan *outlier* atau observasi yang jauh dari distribusi mayoritas data. Keberadaan outlier ini sebelumnya juga terkonfirmasi melalui *violin plot* yang menunjukkan distribusi *right-skewed*, di mana sebagian besar nilai produksi rendah namun terdapat sejumlah kecil data dengan nilai sangat tinggi. Fluktuasi tersebut mulai berkurang pada pembagian data 80:20 dan paling minim pada 90:10 (Gambar b dan c). Hal ini menunjukkan bahwa peningkatan proporsi data latih secara signifikan meningkatkan kemampuan model dalam mengenali pola data. Kurva prediksi pada Gambar 3 terlihat lebih halus dan mendekati kurva aktual, menandakan bahwa model memiliki kemampuan generalisasi yang lebih baik ketika diberi data pelatihan yang lebih banyak. Selanjutnya, dari aspek jumlah decision tree (*n_estimators*), plot dengan 50 pohon (Gambar d) memperlihatkan prediksi yang masih kasar dan fluktuatif. Kurva prediksi sering kali melenceng dari nilai aktual, terutama pada rentang produksi rendah. Seiring peningkatan jumlah pohon menjadi 100, 150, dan 200 (Gambar c, e, dan f), prediksi model tampak semakin stabil dan mendekati nilai aktual, terutama pada rentang nilai menengah hingga rendah. Hal ini menunjukkan bahwa penambahan jumlah pohon berkontribusi pada peningkatan akurasi prediksi karena membantu mengurangi variansi model, konsisten dengan teori dasar Random Forest sebagai metode ensemble. Namun demikian, meskipun tren prediksi sudah mengikuti pola umum, beberapa fluktuasi tetap muncul. Hal ini mengindikasikan bahwa meskipun model sudah cukup kuat, masih terdapat faktor eksternal atau variabel penting lain yang tidak tercakup dalam dataset.

3.4 Evaluasi

Kinerja dari Random Forest Regressor dievaluasi dan dianalisis menggunakan dua metrik utama, yakni Mean Squared Error (MSE) dan R-squared (R^2). Evaluasi memiliki peran krusial dalam menilai sejauh mana model mampu menangkap pola hubungan antara variabel pada dataset dan memastikan model memeiliki efektivitas tinggi dan prediksi yang akurat. Tabel 3 dan 4 merupakan evaluasi model random forest, yang terdiri dari model dasar dan model yang sudah dilakukan fine tuning terhadap pembagian data latih atau jumlah *decision tree*.

Tabel 3. Evaluasi Fine-Tune Pembagian Data Uji

Pembagian Data Uji	MSE	R^2
70:30	0.0040	0.9397

ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

80:20	0.0041	0.8993
90:10	0,0006	0.9887

Tabel 4. Evaluasi Fine-Tune Jumlah Decision Tree

Jumlah Decision Tree	MSE	R^2
50	0,0010	0.9792
100	0,0006	0.9887
150	0.0005	0.9902
200	0.0004	0.9918

Tabel evaluasi performa model Random Forest menunjukkan dua aspek penting yang mempengaruhi akurasi prediksi, yaitu proporsi pembagian data latih dan jumlah decision tree (decision tree) yang digunakan dalam proses pelatihan. Berdasarkan hasil evaluasi terhadap pembagian data, diketahui bahwa semakin besar proporsi data latih yang digunakan, maka akurasi model cenderung meningkat. Hal ini dapat dilihat dari hasil pembagian 90:10 yang menghasilkan nilai Mean Squared Error (MSE) paling rendah, yaitu 0.0006, dan koefisien determinasi (R^2) tertinggi sebesar 0.9887. Nilai tersebut menunjukkan bahwa sebagian besar variasi pada data target dapat dijelaskan oleh model ketika diberi lebih banyak data untuk proses pelatihan. Sebaliknya, pada pembagian 80:20 dan 70:30, performa model sedikit menurun, yang tercermin dari meningkatnya nilai MSE dan turunnya nilai R^2. Hal ini mengindikasikan bahwa dengan lebih sedikit data latih, model memiliki keterbatasan dalam mengenali pola-pola yang kompleks dalam data. Sementara itu, peningkatan jumlah decision tree dalam algoritma Random Forest juga memberikan dampak positif terhadap performa model. Dengan menambah jumlah pohon dari 50 hingga 200, nilai MSE terus mengalami penurunan dari 0.0010 menjadi 0.0004, sedangkan nilai R^2 meningkat dari 0.9792 menjadi 0.9918. Ini menunjukkan bahwa semakin banyak pohon yang digunakan, maka model semakin mampu mengurangi variansi prediksi dan menghasilkan estimasi yang lebih stabil dan akurat.

4. KESIMPULAN

Penelitian ini berhasil membangun model prediksi produksi padi di Indonesia menggunakan algoritma Random Forest Regressor dengan memanfaatkan data agronomis dan iklim selama periode 2018–2024. Proses dimulai dari prapemrosesan data, eksplorasi hubungan antar variabel, hingga pelatihan dan evaluasi model. Hasil evaluasi menunjukkan bahwa model dengan konfigurasi 200 pohon keputusan dan proporsi data latih 90% menghasilkan performa terbaik dengan MSE sebesar 0.0004 dan R^2 sebesar 0.9918. Hal ini menegaskan bahwa Random Forest mampu menangkap kompleksitas hubungan antara variabel input dan target secara akurat. Luas panen dan produktivitas terbukti sebagai prediktor utama, sementara variabel iklim memiliki pengaruh yang lebih rendah secara linier. Meskipun demikian, distribusi data yang tidak merata dan adanya outlier tetap menjadi tantangan, yang ditunjukkan melalui visualisasi violin plot dan fluktuasi hasil prediksi. Penelitian ini masih memiliki keterbatasan dalam cakupan variabel dan kurang mempertimbangkan faktor sosial-ekonomi atau teknologi budidaya. Ke depan, penggabungan variabel tambahan serta eksplorasi metode lain seperti XGBoost atau LSTM diharapkan dapat meningkatkan akurasi dan generalisasi model untuk prediksi produksi padi di berbagai kondisi.

REFERENCES

[1] J. Hutahaean, D. Yusup, And P. Purwantoro, "Perbandingan Metode Linear Regression, Random Forest & K-Nearest Neighbor Untuk Prediksi Produksi Hasil Panen Padi Di Provinsi Jawa Barat," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 8, No. 3, Pp. 3895–3900, 2024.

[2] A. R. Masdian, N. Bashit, And F. Hadi, "Analisis Produktivitas Padi Menggunakan Algoritma Machine Learning Random Forest Di Kabupaten Batang Tahun 2018-2022," *Elipsoida J. Geod. Dan Geomatika*, Vol. 6, No. 1, Pp. 43–51, 2023.

[3] M. A. Musababa, "Implementasi Algoritma Linear Regression Untuk Prediksi Produksi Tanaman Padi Di Kabupaten Grobogan," *Data Sci. Indones.*, Vol. 3, No. 2, Pp. 1–11, 2023.

[4] R. Faizal, A. Abdullah, And M. W. Pangestika, "Perbandingan Random Forest Regressor Dan Decision Tree Regressor Untuk Prediksi Hasil Panen," *J. Coscitech (Computer Sci. Inf. Technol.*, Vol. 6, No. 2, Pp. 247–253, 2025.

[5] N. K. A. Mita, M. F. Siddiq, A. Laurnt, R. Erviana, And R. Kurniawan, "Optimalisasi Ketahanan Pangan: Perbandingan Metode Machine Learning Dan Time Series Dalam Memprediksi Produksi Padi Di Jawa Tengah," In *Prosiding Seminar Nasional Sains Data*, 2024, Pp. 140–153.

[6] Baskoro, Sriyanto, And L. Setya Rini, "Prediksi Penerima Beasiswa Dengan Menggunakan Teknik Data Mining Di Universitas Muhammadiyah Pringsewu," *Semin. Nas. Has. Penelit. Dan Pengabdi. Masy. Inst. Inform. Dan Bisnis Darmajaya*, Vol. 1, No. 2, Pp. 87–94, 2021.

[7] N. Denada, P. Isyanto, And N. Sumarni, "Optimalisasi Media Sosial Tiktok Sebagai Sarana Promosi Di Oculus Photo Studio Cabang Galuh Mas Karawang," *Manag. Stud. Entrep. J.*, Vol. 4, No. 6, Pp. 10070–10085, 2023.

[8] E. Fitri And S. N. Nugraha, "Optimasi Kinerja Linear Regression, Random Forest Regression Dan Multilayer Perceptron Pada Prediksi Hasil Panen," *Inti Nusa Mandiri*, Vol. 18, No. 2, Pp. 210–217, 2024.

[9] S. Sobari, A. I. Purnamasari, A. Bahtiar, And K. Kaslani, "Meningkatkan Model Prediksi Kelulusan Santri Tahfidz Di Pondok Pesantren Al-Kautsar Menggunakan Algoritma Random Forest," *J. Inform. Dan Tek. Elektro Terap.*, Vol. 13, No. 1, 2025.

ISSN 2808-375X (Media Online)
Vol 4, No 3, July 2025
Hal 337−345
DOI 10.47065/comforch.v4i3.2125

[10] N. Hadi And J. Benedict, "Implementasi Machine Learning Untuk Prediksi Harga Rumah Menggunakan Algoritma Random Forest," *Comput. J. Comput. Sci. Inf. Syst.*, Vol. 8, No. 1, Pp. 50–61, 2024.
[11] S. A. Putri, N. Selayanti, M. Kristanaya, M. P. Azzahra, M. G. Navsih, And K. M. Hindrayani, "Penerapan Machine Learning Algoritma Random Forest Untuk Prediksi Penyakit Jantung," In *Prosiding Seminar Nasional Sains Data*, 2024, Pp. 895–906.
[12] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," *J. Infomedia Tek. Inform. Multimedia, Dan Jar.*, Vol. 9, No. 1, Pp. 16–21, 2024.
[13] H. Sunaryanto, M. A. Hasan, And G. Guntoro, "Classification Analysis Of Unilak Informatics Engineering Students Using Support Vector Machine (Svm), Iterative Dichotomiser 3 (Id3), Random Forest And K-Nearest Neighbors (Knn)," *It J. Res. Dev.*, Vol. 7, No. 1, Pp. 36–42, 2022, Doi: 10.25299/Itjrd.2022.8912.
[14] I. L. Mulyahati, "Implementasi Machine Learning Prediksi Harga Sewa Apartemen Menggunakan Algoritma Random Forest Melalui Framework Website Flask Python (Studi Kasus: Apartemen Di Dki Jakarta Pada Website Mamikos. Com)," 2020.
[15] N. Maulidah, M. Maulidah, R. Supriyadi, H. Nalatissifa, S. Diantika, And A. Fauzi, "Prediksi Kualitas Air Menggunakan Metode Random Forest, Decision Tree, Dan Gradient Boosting," *J. Khatulistiwa Inform.*, Vol. 12, No. 1, Pp. 1–6, 2024, Doi: 10.31294/Jki.V12i1.16004.
[16] I. Padiku And A. Lahinta, "Penerapan Clustering K-Means Untuk Mendukung Pengelolaan Koleksi Pada Perpustakaan Fakultas Teknik Universitas Negeri Gorontalo," *J. Tek.*, Vol. 20, No. 1, Pp. 54–62, 2022.
[17] Y. Wendra, A. Alwendi, A. Ardi, And D. Aldo, "Metode Case Based Reasoning Untuk Identifikasi Penyakit Tanaman Padi," *Jursima (Jurnal Sist. Inf. Dan Manajemen)*, Vol. 8, No. 2, Pp. 103–110, 2020.
[18] A. Mulyanto And E. Apriyanti, "Penerapan Sistem Informasi Penanggulangan Hama Padi Dengan Menggunakan Algoritma K-Nearest Neighbor Dan Metode Case Based Reasoning (Studi Kasus Kecamatan Kedung Waringin Kabupaten Bekasi)," *J. Inform. Simantik*, Vol. 7, No. 1, Pp. 1–5, 2022.