JURNAL JTIK
(JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI)
KITA Komunitas Informasi Teknologi Aceh
OPEN ACCESS
E-ISSN: 2580-1643
Volume 11, Nomor 1, Januari 2027 | Pages: 160-171 | DOI: https://doi.org/10.35870/jtik.v11i1.7567

# Prediksi Hasil Tanaman Padi Menggunakan *Random Forest* dengan Seleksi Fitur Berbasis *Feature Importance* dan Optimasi *Hyperparameter GridSearchCV*

Steven Andrian Pranata Wicaksono [1*], Mutaqin Akbar [2]

[1*,2] Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Mercu Buana Yogyakarta, Kabupaten Bantul, Provinsi Daerah Istimewa Yogyakarta, Indonesia.

*Corresponding Email*: steven.andrian889@gmail.com [1*]

*Article History: Received: 15 June 2026, Revision: 8 July 2026, Accepted: 14 July 2026, Available Online: 1 January 2027.*

## Abstract

Rice production in Indonesia fluctuates due to agronomic, spatial, and climatic factors. This study develops a rice production validation model using Random Forest Regressor with feature importance-based feature selection and GridSearchCV hyperparameter optimization. Agricultural data from BPS for 2018-2024 were combined with annual NASA POWER weather data from 37 provinces in Indonesia. The model predicts rice productivity and converts the prediction into production using actual harvested area. Feature selection reduced 32 initial predictors to 16 final features. The optimized model achieved production $R^2$ of 99.79%, adjusted $R^2$ of 99.73%, RMSE of 115,600.28 tons, MAE of 61,194.75 tons, MAPE of 6.95%, and SMAPE of 6.86%. It is important to note that the high production $R^2$ of 99.79% is substantially driven by the mathematical dominance of actual harvested area as a multiplier in the production conversion formula, rather than solely reflecting the predictive power of the climate model. The core predictive performance of the model at the productivity level ($R^2$ = 75.45%, MAPE = 6.95%) more accurately represents the model's generalization capability. Research limitations include limited historical data for newly established provinces in Papua and the relatively short study period of 2018-2024. These results indicate that the proposed method provides accurate validation for rice production analysis.

**Keyword:** Feature Importance; GridSearchCV; NASA POWER; Rice Production; Random Forest.

## Abstrak

Produksi padi di Indonesia berfluktuasi karena dipengaruhi faktor agronomis, spasial, dan iklim. Penelitian ini membangun model validasi produksi padi menggunakan Random Forest Regressor dengan seleksi fitur berbasis feature importance dan optimasi hyperparameter GridSearchCV. Data pertanian BPS periode 2018-2024 digabungkan dengan data cuaca tahunan NASA POWER dari 37 provinsi di Indonesia. Model memprediksi produktivitas padi, kemudian hasil prediksi dikonversi menjadi produksi menggunakan luas panen aktual. Seleksi fitur mereduksi 32 prediktor awal menjadi 16 fitur final. Model teroptimasi memperoleh $R^2$ produksi sebesar 99,79%, adjusted $R^2$ sebesar 99,73%, RMSE sebesar 115.600,28 ton, MAE sebesar 61.194,75 ton, MAPE sebesar 6,95%, dan SMAPE sebesar 6,86%. Perlu dicatat bahwa nilai $R^2$ produksi sebesar 99,79% dipengaruhi secara signifikan oleh dominasi matematis variabel luas panen aktual sebagai pengali dalam formula konversi produksi, bukan semata-mata mencerminkan keandalan prediksi iklim model. Performa prediksi inti model pada tingkat produktivitas ($R^2$ = 75,45%, MAPE = 6,95%) lebih mencerminkan kemampuan generalisasi model yang sesungguhnya. Keterbatasan penelitian mencakup minimnya data historis untuk provinsi pemekaran Papua dan periode studi yang relatif singkat (2018-2024). Hasil ini menunjukkan bahwa metode yang diusulkan mampu memberikan validasi produksi padi yang akurat.

**Kata Kunci:** Feature Importance; GridSearchCV; NASA POWER; Produksi Padi; Random Forest.

E-ISSN: 2580-1643

# 1. Pendahuluan

Indonesia merupakan negara dengan tingkat konsumsi beras tertinggi di Asia Tenggara, di mana padi menjadi komoditas pangan strategis bagi lebih dari 270 juta penduduk dengan kebutuhan konsumsi beras nasional mencapai sekitar 30 juta ton per tahun. Data Badan Pusat Statistik (BPS) mencatat bahwa produksi padi nasional mengalami penurunan signifikan pada tahun 2019 sebesar 7,75 juta ton atau turun sekitar 7,6% dibandingkan tahun 2018, dari 59,20 juta ton menjadi hanya 54,60 juta ton, sehingga berdampak pada ketersediaan stok beras nasional dan memicu kenaikan harga yang berdampak langsung pada daya beli masyarakat (Manurung *et al.*, 2025). Fluktuasi produksi ini dipengaruhi oleh interaksi kompleks antara faktor agronomis seperti luas panen dan produktivitas lahan, serta faktor iklim yang meliputi curah hujan, suhu udara, kelembaban relatif, dan tekanan udara di setiap provinsi (Adin Musababa, 2024). Oleh karena itu, model prediksi yang mampu menangkap hubungan non-linier antarvariabel tersebut sangat dibutuhkan sebagai alat bantu pengambilan keputusan bagi petani, pemerintah, dan pemangku kepentingan sektor pertanian dalam menjaga stabilitas ketahanan pangan nasional.

Berbagai metode prediksi konvensional telah diterapkan untuk memodelkan produksi padi. Regresi linier sederhana mampu menjelaskan sekitar 63-82% variasi produksi menggunakan variabel luas lahan dan kondisi iklim (Adin Musababa, 2024), sementara regresi linier berganda yang diterapkan pada data produksi padi di Jawa Barat menghasilkan $R^2$ = 0,79 (Kharisma S *et al.*, 2025). Meskipun memberikan estimasi awal yang memadai, kedua metode ini memiliki keterbatasan fundamental karena mengasumsikan hubungan linier antarvariabel dan tidak mampu menangkap interaksi non-linier yang kompleks antara faktor iklim dan agronomis (Faizal *et al.*, 2025). *Random Forest* dipilih sebagai metode prediksi karena unggul dalam menangani data kompleks dengan beragam tipe fitur kategorikal, diskrit, dan kontinu. Algoritma ini merupakan ensemble pohon keputusan acak yang akurat dan tahan terhadap overfitting (Breiman, 2001; Selayanti *et al.*, 2024), sekaligus menyediakan estimasi *feature importance* secara inheren untuk mengidentifikasi variabel yang paling berpengaruh (Tantyoko *et al.*, 2023). *Random Forest* juga terbukti menghasilkan akurasi tertinggi dibandingkan *SVM*, *Gradient Boosting*, dan *Decision Tree* dalam prediksi cuaca (Risanti & Suhendar, 2024), memperkuat relevansinya untuk data yang mengandung banyak variabel iklim. Meskipun demikian, konfigurasi hyperparameter perlu dioptimalkan karena memengaruhi kinerja secara signifikan (Breiman, 2001). Beberapa penelitian terdahulu telah menerapkan *Random Forest* untuk prediksi produksi padi; (Manurung *et al.*, 2025) memperoleh MSE = 0,0004 dan $R^2$ = 0,9918 menggunakan data BPS dan *NASA POWER* namun tanpa seleksi fitur maupun tuning hyperparameter sistematis; (Jati *et al.*, 2025) menerapkan *GridSearchCV* dan memperoleh $R^2$ = 0,986 namun tanpa seleksi fitur; serta (Masdian *et al.*, 2023) dan (Hutahaean *et al.*, 2024) memiliki cakupan wilayah terbatas sehingga tidak merepresentasikan keragaman kondisi pertanian nasional. Pada data pertanian yang memiliki banyak variabel iklim, seleksi fitur perlu dilakukan untuk memperbaiki kualitas model karena fitur yang tidak relevan dapat meningkatkan risiko overfitting serta menurunkan interpretabilitas model. (Tantyoko *et al.*, 2023) membuktikan bahwa seleksi fitur berbasis *feature importance* pada *Random Forest* meningkatkan F1-score dari 87,21% menjadi 92,23%. Selain itu, optimasi hyperparameter diperlukan untuk memastikan konfigurasi model yang paling optimal. (Gori & Hestiningtyas, 2024) menunjukkan bahwa kombinasi seleksi fitur dan *GridSearchCV* meningkatkan akurasi *Random Forest* dari 88,04% menjadi 91,85%, membuktikan bahwa kedua teknik tersebut memiliki efek yang saling melengkapi.

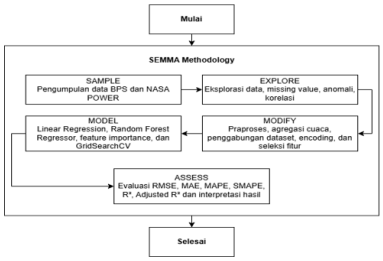
(Maisat Eka Darmawan & Fauzan Dianta, 2023) juga membuktikan bahwa *GridSearchCV* meningkatkan akurasi model dari 83% menjadi 86% melalui pencarian sistematis terhadap kombinasi hyperparameter terbaik. Berdasarkan penelitian-penelitian tersebut, terdapat tiga kesenjangan yang perlu diisi. Pertama, belum ada penelitian yang mengintegrasikan seleksi fitur berbasis *feature importance RF* dengan optimasi *GridSearchCV* secara bersamaan untuk prediksi produksi padi di Indonesia. Kedua, sebagian besar penelitian terdahulu menggunakan data BMKG yang umumnya terbatas pada sekitar 10 variabel iklim dan memiliki cakupan historis yang lebih pendek, sedangkan *NASA POWER* menyediakan hingga 23 parameter iklim, termasuk radiasi matahari dan indeks UV, yang dapat diakses gratis melalui API resmi dengan cakupan seluruh provinsi, termasuk wilayah terpencil. Kelebihan *NASA POWER* terletak pada konsistensi spasial-temporal berbasis satelit dan model reanalisis atmosfer, sehingga lebih sesuai untuk pemodelan skala nasional yang mencakup 37 provinsi. Ketiga, cakupan wilayah penelitian sebelumnya terbatas pada satu kabupaten atau provinsi (Hutahaean *et al.*, 2024; Masdian *et al.*, 2023) sehingga tidak merepresentasikan keragaman kondisi pertanian nasional. Kebaruan penelitian ini terletak pada integrasi simultan seleksi fitur berbasis *feature importance* dan optimasi *GridSearchCV* dengan

E-ISSN: 2580-1643

skema *TimeSeriesSplit* pada data 37 provinsi Indonesia periode 2018-2024 yang bersumber dari BPS dan *NASA POWER API*, dengan tujuan menghasilkan model prediksi produksi padi yang lebih akurat, stabil, dan dapat digeneralisasi untuk mendukung pengambilan keputusan di sektor pertanian nasional.

## 2. Metode Penelitian

Alur penelitian menggunakan metodologi Sample, Explore, Modify, Model, Assess (SEMMA) yang dipilih karena sistematis, terstruktur, dan telah terbukti efektif pada berbagai penelitian data mining di berbagai domain (Nur Fauzi *et al.*, 2025). Tahapan penelitian ditunjukkan pada Gambar 1.




Gambar 1. Alur Metodologi SEMMA

### *Sample*

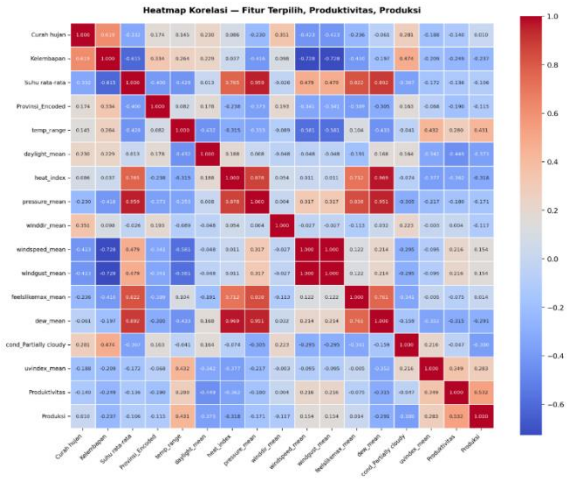
Data yang digunakan terdiri dari dua sumber sekunder yang dipilih berdasarkan ketersediaan, kelengkapan variabel, dan aksesibilitas publik. Pertama, data pertanian dari BPS periode 2018-2024 mencakup luas panen, produktivitas, dan produksi padi pada 37 provinsi di Indonesia. Kedua, data cuaca dari API NASA POWER berdasarkan koordinat pusat masing-masing provinsi, mencakup lebih dari beberapa parameter iklim meliputi suhu, curah hujan, kelembaban, tekanan udara, kecepatan angin, radiasi matahari, dan indeks UV. API resmi NASA POWER dipilih karena menyediakan data dalam format JSON yang terstruktur dan lebih stabil dibandingkan metode web scraping, yang rentan terhadap perubahan struktur halaman, mekanisme CAPTCHA, dan deteksi bot (Rizquina & Ratnasari, 2023). Data cuaca harian diagregasi menjadi data tahunan: curah hujan dijumlahkan sebagai total tahunan karena bersifat kumulatif, sedangkan parameter berkelanjutan dihitung sebagai rata-rata tahunan (Manurung *et al.*, 2025). Kedua sumber digabungkan berdasarkan provinsi dan tahun menghasilkan dataset akhir sebanyak 239 baris dan 32 fitur. Variabel target utama adalah produktivitas padi (ku/ha) yang dikonversi menjadi produksi (ton) menggunakan rumus (Manurung *et al.*, 2025):

$$Produksi\ (ton) = \frac{Produktivitas\ (ku/ha) \times Luas\ Panen\ (ha)}{10}$$

Pembagian data dilakukan secara kronologis untuk menjaga sifat temporal dan mencegah data leakage: data 2018-2022 sebagai data latih (165 baris, 69,04%) dan data 2023-2024 sebagai data uji (74 baris, 30,96%).

### *Explore*

Tahap eksplorasi mencakup statistik deskriptif, deteksi missing value dan anomali, serta korelasi Pearson; tidak ditemukan missing value atau anomali sehingga seluruh 239 baris digunakan. Empat provinsi pemekaran Papua (Papua Barat Daya, Papua Pegunungan, Papua Selatan, Papua Tengah) hanya memiliki data 2023-2024, sehingga ditambahkan fitur biner Provinsi_Baru untuk menandai riwayat data terbatas. Kesenjangan spasial ini berpotensi menimbulkan bias akurasi sektoral karena model sulit mempelajari pola historis provinsi pemekaran dibandingkan provinsi Pulau Jawa (2018-2024). Untuk mengurangi bias, pembagian data dilakukan secara kronologis agar setiap provinsi terwakili sesuai periode tersedia tanpa oversampling atau interpolasi; evaluasi akurasi per provinsi tetap dilakukan untuk mendeteksi bias. Analisis korelasi menunjukkan produktivitas produksi 0,532 (positif), temp_range produksi 0,431 (positif), dan daylight_mean–produktivitas -0,449 (negatif). Korelasi sedang ini mengindikasikan kemungkinan hubungan non-linier antarvariabel, mendukung penggunaan Random Forest. Heatmap korelasi disajikan di Gambar 2.




Gambar 2. Heatmap Korelasi Fitur

### *Modify*

Praproses data dilakukan melalui lima langkah sistematis. Pertama, konversi format angka Indonesia (titik sebagai pemisah ribuan, koma sebagai pemisah desimal) ke format numerik Python. Kedua, normalisasi nama provinsi menggunakan pemetaan untuk menghindari

E-ISSN: 2580-1643

kegagalan penggabungan data. Ketiga, pembuatan fitur turunan meliputi temp_range (selisih suhu maksimum dan minimum), heat_index (perkalian suhu rata-rata dan kelembaban dibagi 100), dan hujan_per_hari (total curah hujan tahunan dibagi 365). Keempat, encoding variabel kategorikal provinsi menggunakan Label Encoding. Kelima, standardisasi seluruh fitur input menggunakan StandardScaler yang di-fit hanya pada data latih kemudian diterapkan pada data uji untuk mencegah data leakage (Jati *et al.*, 2025; Ristyawan *et al.*, 2025).

***Model***

Model RF awal dilatih menggunakan seluruh 32 fitur untuk memperoleh nilai feature importance berdasarkan rata-rata penurunan MSE akibat pemisahan node yang melibatkan setiap fitur di seluruh pohon (Breiman, 2001). Fitur dengan *importance* di bawah threshold 0,005 dieliminasi, kecuali tiga fitur wajib agronomis yaitu curah hujan, kelembaban, dan suhu rata-rata yang tetap dipertahankan karena terbukti berpengaruh terhadap pertumbuhan padi (Kharisma S *et al.*, 2025). *GridSearchCV* mencoba seluruh kombinasi nilai dalam grid secara sistematis dan mengevaluasi setiap kombinasi menggunakan TimeSeriesSplit 5-fold untuk menjaga urutan kronologis dan mencegah kebocoran data (Jati *et al.*, 2025; Maisat Eka Darmawan & Fauzan Dianta, 2023). Metrik optimasi yang digunakan adalah neg_mean_absolute_error. Grid hyperparameter disajikan pada Tabel 1, menghasilkan total 648 kombinasi × 5 fold = 3.240 iterasi pelatihan. Konfigurasi optimal dipilih menggunakan persamaan (Maisat Eka Darmawan & Fauzan Dianta, 2023):

$$\theta^* = \arg\min_{\theta \in \Theta} Error_{CV}(\theta)$$

Dimana θ adalah kombinasi hyperparameter terbaik, Θ adalah ruang kandidat hyperparameter, dan ErrorCV(θ) adalah error rata-rata validasi silang untuk kombinasi θ.

Tabel 1. Grid Hyperparameter GridSearchCV

| Hyperparameter | Kandidat Nilai |
|---|---|
| n_estimators | 100, 200, 300 |
| max_depth | 4, 6, 8, None |
| min_samples_split | 2, 5, 10 |
| min_samples_leaf | 1, 2, 4 |
| min_impurity_decrease | 0,0; 0,0001 |
| max_features | sqrt, 0,3, 0,5 |
| bootstrap | True |

Model final Random Forest Regressor dilatih menggunakan seluruh 165 data latih dengan hyperparameter terbaik hasil GridSearchCV. Sebagai pembanding, model baseline Linear Regression dilatih menggunakan fitur yang sama (Fitri & Nugraha, 2024; Muchtar & Afiyati, 2024). Random Forest bekerja dengan membangun sejumlah pohon keputusan secara paralel menggunakan mekanisme bagging dan keacakan fitur pada setiap node, dengan prediksi akhir dihitung sebagai rata-rata seluruh pohon (Breiman, 2001; Selayanti *et al.*, 2024):

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(X)$$

Dimana ŷ adalah prediksi akhir, B adalah jumlah pohon, dan Tb(X) adalah prediksi pohon ke-b untuk input X.

Seluruh pengolahan dan analisis data dilakukan menggunakan Python 3.12.3. Manipulasi data memanfaatkan Pandas 2.2.3 dan NumPy 2.2.4, pembangunan model menggunakan Scikit-learn 1.6.1, sedangkan visualisasi dibuat dengan Matplotlib 3.10.1 dan Seaborn 0.13.2. Seluruh kode dikembangkan di Visual Studio Code 1.99 pada Windows 11 untuk mendukung reprodusibilitas penelitian.

***Assess***

Evaluasi model dilakukan pada dua tingkat keluaran: evaluasi produktivitas (ku/ha) sebagai output langsung model, dan validasi produksi (ton) setelah produktivitas prediksi dikonversi menggunakan luas panen aktual. Metrik evaluasi yang digunakan meliputi $R^2$, Adjusted $R^2$, RMSE, MAE, MAPE, dan SMAPE (Manurung *et al.*, 2025; Masdian *et al.*, 2023):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

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

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

E-ISSN: 2580-1643

Di mana *yi* adalah nilai aktual, *ŷi* adalah nilai prediksi, $\bar{y}$ adalah rata-rata nilai aktual, *n* adalah jumlah data, dan *k* adalah jumlah fitur. Nilai MAPE di bawah 10% menunjukkan model termasuk kategori akurat. Sebagai analisis tambahan, dibangun empat model klasifikasi bertahap untuk mengkuantifikasi kontribusi individual setiap teknik optimasi terhadap peningkatan performa model, yaitu: RM1 (*Baseline Random Forest* tanpa optimasi), RM2 (*Random Forest* dengan seleksi fitur berbasis *feature importance*), RM3 (Random Forest dengan optimasi *hyperparameter GridSearchCV* tanpa seleksi fitur), dan RM4 (*Random Forest* dengan seleksi fitur dan *GridSearchCV* secara bersamaan).

# 3. Hasil dan Pembahasan

## 3.1 Hasil

Penelitian ini menyajikan hasil pemodelan *Random Forest Regressor* yang dikombinasikan dengan seleksi fitur berbasis feature importance dan optimasi *hyperparameter GridSearchCV* pada data produksi padi 37 provinsi Indonesia periode 2018-2024. Tabel 2 menyajikan ringkasan dataset penelitian.

Tabel 2. Ringkasan Dataset dan Hasil Penelitian

| Komponen | Hasil |
|---|---|
| Jumlah data akhir | 239 baris |
| Jumlah provinsi | 37 provinsi |
| Rentang tahun | 2018–2024 |
| Data latih | 2018–2022 (165 baris; 69,04%) |
| Data uji | 2023–2024 (74 baris; 30,96%) |
| Fitur awal | 32 fitur |
| Fitur final (regresi) | 16 fitur |
| Target utama model | Produktivitas padi (ku/ha) |
| $R^2$ produksi test | 99,79% |
| MAPE produksi test | 6,95% |

***Hasil Seleksi Fitur***

Seleksi fitur berhasil mereduksi 32 fitur awal menjadi 16 fitur final menggunakan nilai *feature importance* internal *Random Forest* dengan ambang batas *threshold* 0,005 (0,5%). Tiga fitur wajib agronomis yaitu curah hujan, kelembaban, dan suhu rata-rata tetap dipertahankan meskipun nilai *importance*-nya di bawah threshold karena memiliki peran langsung terhadap pertumbuhan padi (Kharisma S *et al.*, 2025). Tabel 3 menunjukkan peringkat *feature importance* seluruh 16 fitur final.

Tabel 3. Peringkat Feature Importance 16 Fitur Final

| No | Fitur | Importance (%) |
|---|---|---|
| 1 | Provinsi_Encoded | 61,04 |
| 2 | temp_range | 8,15 |
| 3 | daylight_mean | 7,43 |
| 4 | heat_index | 5,28 |
| 5 | pressure_mean | 3,24 |
| 6 | winddir_mean | 1,85 |
| 7 | windspeed_mean | 1,27 |
| 8 | windgust_mean | 1,24 |
| 9 | feelslikemax_mean | 1,10 |
| 10 | dew_mean | 1,07 |
| 11 | Kelembapan | 0,97 |
| 12 | cond_Partially cloudy | 0,69 |
| 13 | uvindex_mean | 0,59 |
| 14 | feelslike_mean | 0,53 |
| 15 | Suhu rata-rata | 0,47 |
| 16 | Curah hujan | 0,28 |

*Provinsi_Encoded* menjadi fitur paling dominan dengan *importance* sebesar 61,04%, menunjukkan bahwa perbedaan karakteristik antarprovinsi berkontribusi terhadap variasi hasil prediksi model, sejalan dengan temuan bahwa mekanisme *feature importance Random Forest* mampu menghasilkan interpretasi yang bermakna secara substantif lintas domain (Handayani & Qutub, 2025). Namun, dominasi fitur ini perlu ditafsirkan secara hati-hati karena *Provinsi_Encoded*

E-ISSN: 2580-1643

merepresentasikan perbedaan karakteristik spasial, bukan faktor agronomis langsung. Fitur iklim seperti temp_range (8,15%), daylight_mean (7,43%), dan heat_index (5,28%) menempati posisi kedua hingga keempat, mengkonfirmasi pentingnya informasi iklim terhadap produktivitas padi (Manurung *et al.*, 2025).

Tabel 4. CV $R^2$ Berdasarkan Jumlah Fitur

| Jumlah Fitur | CV $R^2$ | Jumlah Fitur | CV$R^2$ |
|---|---|---|---|
| 2 | 0,7952 | 11 | 0,8145 |
| 3 | 0,7912 | 12 | 0,8080 |
| 4 | 0,8269 | 13 | 0,8061 |
| 5 (tertinggi) | 0,8283 | 14 | 0,8034 |
| 6 | 0,8226 | 15 | 0,8050 |
| 7 | 0,8168 | 16 (final) | 0,8020 |
| 8 | 0,8156 | 17 | 0,7911 |
| 9 | 0,8208 | 18 | 0,7845 |
| 10 | 0,8202 | 19 | 0,7835 |

Meskipun nilai CV $R^2$ tertinggi diperoleh pada 5 fitur (CV $R^2$ = 0,8283), model akhir tetap menggunakan 16 fitur final. Tabel 4 menunjukkan nilai CV $R^2$ untuk setiap jumlah fitur dari 2 hingga 19, yang menjadi dasar pemilihan jumlah fitur optimal. Keputusan ini diambil karena beberapa pertimbangan: (1) stabilitas model penambahan fitur ke-6 hingga ke-16 hanya menurunkan CV $R^2$ secara bertahap dari 0,8283 ke 0,8020, bukan penurunan drastis; (2) kelengkapan informasi iklim fitur seperti kelembaban, suhu rata-rata, dan curah hujan memiliki makna agronomis langsung terhadap pertumbuhan padi meskipun nilai importance-nya kecil (Kharisma S *et al.*, 2025); dan (3) keterkaitan domain eliminasi variabel iklim utama hanya berdasarkan angka importance akan mengurangi kemampuan model dalam konteks pertanian. Temuan ini sejalan dengan (Tantyoko *et al.*, 2023) yang menyatakan bahwa seleksi fitur meningkatkan F1-score dari 87,21% menjadi 92,23% melalui eliminasi fitur tidak informatif.

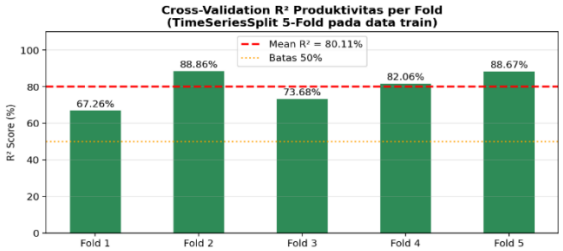
### Hasil Optimasi *Hyperparameter*

*GridSearchCV* dengan *TimeSeriesSplit* 5-fold menghasilkan 3.240 pengulangan pelatihan dari 648 kombinasi hyperparameter. Tabel 5 menunjukkan konfigurasi hyperparameter terbaik yang diperoleh dengan Best CV MAE produktivitas sebesar 2,9347 ku/ha.

Tabel 5. Konfigurasi Hyperparameter Terbaik

| Parameter | Nilai Terbaik |
|---|---|
| n_estimators | 100 |
| max_depth | None |
| max_features | 0,5 |
| min_impurity_decrease | 0,0001 |
| min_samples_leaf | 1 |
| min_samples_split | 2 |
| bootstrap | True |
| Best CV MAE Produktivitas | 2,9347 ku/ha |

Konfigurasi terbaik menunjukkan bahwa model tidak membutuhkan pembatasan max_depth, tetapi tetap menggunakan min_impurity_decrease sebesar 0,0001 agar pemecahan node yang tidak informatif dapat dikontrol. Evaluasi cross-validation menggunakan TimeSeriesSplit 5-fold menunjukkan rata-rata $R^2$ sebesar 80,11% dengan rentang 67,26%-88,86%, sebagaimana ditunjukkan pada Gambar 3. Sebagai pembanding, KFold 5-fold memperoleh rata-rata $R^2$ sebesar 83,03% ± 5,66% dengan MAE rata-rata 2,6013 ku/ha. Nilai KFold yang sedikit lebih tinggi menunjukkan model lebih kuat menangkap pola variasi karakterristik antar provinsi, sedangkan TimeSeriesSplit memberikan pengujian yang lebih ketat karena mempertahankan urutan waktu (Maisat Eka Darmawan & Fauzan Dianta, 2023).




Gambar 3. CV $R^2$ per Fold (TimeSeriesSplit 5-Fold)

E-ISSN: 2580-1643

**Evaluasi Model Regresi dan Validasi**

Evaluasi dilakukan pada dua tingkat keluaran: produktivitas (ku/ha) sebagai output langsung model, dan produksi (ton) setelah konversi menggunakan luas panen aktual. Tabel 6 dan Tabel 7 menyajikan ringkasan metrik evaluasi pada data latih dan data uji.

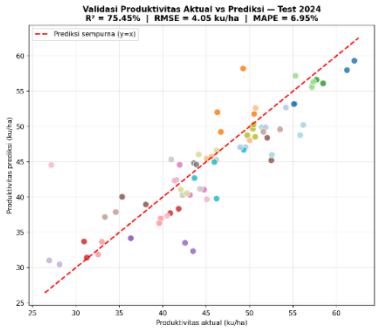
Tabel 6. Metrik $R^2$ dan Adjusted $R^2$

| Data | $R^2$ | Adj. $R^2$ |
|---|---|---|
| Produktivitas Train | 97,46% | - |
| Produktivitas Test | 75,45% | - |
| Produksi Train | 99,98% | - |
| Produksi Test | 99,79% | 99,73% |

Tabel 7. Metrik RMSE, MAE, MAPE, dan SMAPE

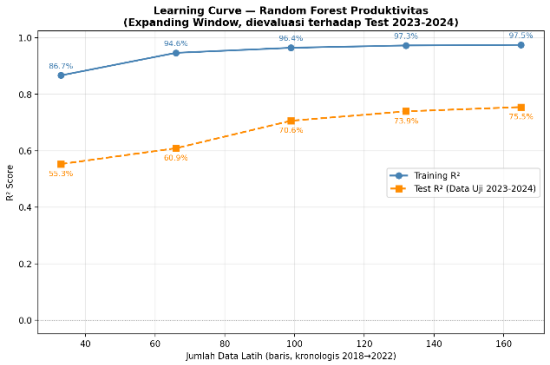
| RMSE | MAE | MAPE | SMAPE |
|---|---|---|---|
| 1,41 ku/ha | 0,96 ku/ha | 2,26% | 2,24% |
| 4,05 ku/ha | 2,96 ku/ha | 6,95% | 6,86% |
| 35.117 ton | 18.517 ton | 2,26% | 2,24% |
| 115.600 ton | 61.195 ton | 6,95% | 6,86% |

Pada data uji, model memperoleh $R^2$ produktivitas sebesar 75,45% dengan MAPE sebesar 6,95%, menunjukkan kesalahan relatif di bawah 10% sehingga masuk kategori akurat. Setelah konversi ke produksi menggunakan luas panen aktual, $R^2$ meningkat signifikan menjadi 99,79% dengan Adjusted $R^2$ = 99,73%, karena luas panen aktual menjadi pengali yang mendominasi skala produksi sementara produktivitas prediksi hanya berperan menentukan efisiensi lahan. Perbandingan dengan *baseline Linear Regression* menunjukkan peningkatan performa yang signifikan: Random Forest memperoleh RMSE = 115.600,28 ton, MAE = 61.194,75 ton, dan MAPE = 6,95%, jauh lebih baik dibandingkan Linear Regression dengan RMSE = 512.712,60 ton, MAE = 228.753,55 ton, dan MAPE = 15,63%. Penurunan RMSE sebesar 77,44% membuktikan keunggulan Random Forest dalam menangkap hubungan non-linier antara variabel iklim, spasial, dan produktivitas padi, sejalan dengan temuan bahwa RF secara konsisten mengungguli Linear Regression baik dari sisi $R^2$ maupun MAE pada data tabular pertanian (Fitri & Nugraha, 2024; Muchtar & Afiyati, 2024). Selisih agregat produksi aktual (106.695.691,58 ton) dan prediksi (103.329.102,00 ton) hanya sebesar 3,16%, tergolong kecil dalam skala nasional.




Gambar 4. Produktivitas Aktual vs Prediksi Data Uji

Gambar 4 Validasi Produktivitas Aktual vs Prediksi pada Data Uji 2023-2024. Setiap titik mewakili satu observasi provinsi-tahun (n=74). Garis merah putus-putus menunjukkan garis prediksi sempurna (y = x). Dispersi titik di sekitar garis sempurna mencerminkan kesalahan prediksi model pada tingkat produktivitas ($R^2$ = 75,45%; RMSE = 4,05 ku/ha; MAPE = 6,95%), yang lebih rendah dari $R^2$ produksi (99,79%) karena tidak melibatkan pengali luas panen actual.

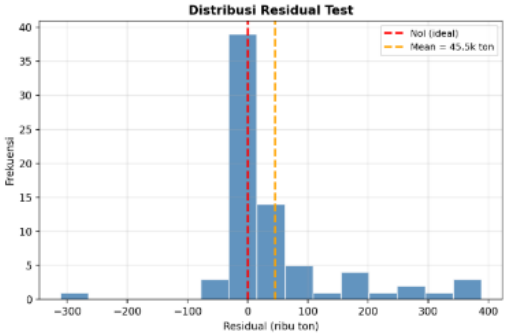


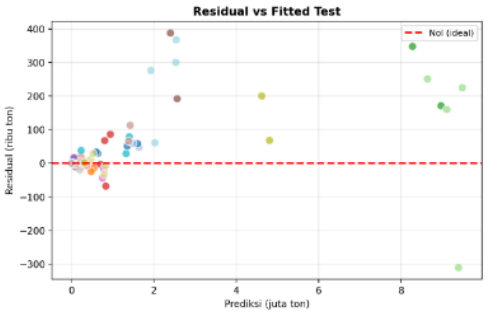

Gambar 5. *Learning Curve Random Forest*

Gambar 5 menampilkan learning curve Random Forest pada level produktivitas dengan pendekatan expanding window, yaitu model dilatih ulang secara bertahap menggunakan data latih kronologis 2018-2022 (20% hingga 100%) dan dievaluasi pada data uji 2023-2024. Training $R^2$ cepat stabil di kisaran 96-97% ketika data latih mencapai sekitar 100 baris, sedangkan Test $R^2$ naik dari 55,3% menjadi 75,5% pada 165 baris data latih. Pola ini menunjukkan model masih berpotensi meningkat jika tersedia data historis yang lebih panjang, sejalan dengan keterbatasan periode studi 2018-2024. Analisis residual menunjukkan median residual mendekati nol (-5.534,50 ton), mengindikasikan tidak ada systematic bias yang signifikan meskipun model

E-ISSN: 2580-1643

cenderung sedikit mengestimasi lebih rendah secara agregat (Masdian *et al.*, 2023), seperti diperlihatkan pada Gambar 6.




Gambar 6. Analisis Residual Produksi

**Analisis Distribusi akurasi per Provinsi**

Distribusi akurasi dari 74 data uji periode 2023-2024 ditunjukkan pada Tabel 8. Sebanyak 79,73% data uji berada dalam kategori akurat atau sangat akurat.

Tabel 8. Distribusi Akurasi per Kategori Error

| Status | Kriteria | Jumlah | Persen |
|---|---|---|---|
| Sangat Akurat | Error < 5% | 39 | 52,70% |
| Akurat | Error 5–10% | 20 | 27,03% |
| Perlu Perhatian | Error > 10% | 15 | 20,27% |
| Total | - | 74 | 100% |

Provinsi dengan error rata-rata tertinggi adalah Papua Barat Daya (33,16%), Kep. Bangka Belitung (23,48%), dan DKI Jakarta (15,29%). Error tinggi pada Papua Barat Daya disebabkan keterbatasan riwayat data karena provinsi ini merupakan hasil pemekaran yang hanya tersedia pada periode pengujian, sehingga model tidak memiliki histori provinsi tersebut dalam data latih dan prediksi lebih bergantung pada pola umum provinsi lain. DKI Jakarta dan Kep. Bangka Belitung memiliki skala produksi yang sangat kecil dibandingkan provinsi lain sehingga fluktuasi kecil secara absolut menghasilkan error relatif yang besar. Tabel 9 menunjukkan rata-rata error per provinsi dan sepuluh provinsi dengan rata-rata error tertinggi.

Tabel 9. Provinsi dengan Error Tertinggi

| Provinsi | Rerata Error (%) | Jumlah Data |
|---|---|---|
| Papua Barat Daya | 33,16 | 2 (2023-2024) |
| Kep. Bangka Belitung | 23,48 | 2 (2023-2024) |
| DKI Jakarta | 15,29 | 2 (2023-2024) |
| Kep. Riau | 11,71 | 2 (2023-2024) |
| Sumatera Selatan | 11,62 | 2 (2023-2024) |
| Maluku Utara | 10,48 | 2 (2023-2024) |
| Lampung | 10,44 | 2 (2023-2024) |
| Papua Barat | 9,90 | 2 (2023-2024) |
| Sulawesi Utara | 8,06 | 2 (2023-2024) |

**Evaluasi Klasifikasi Tambahan**

Sebagai analisis pendukung, dibangun empat model klasifikasi bertahap untuk mengukur kontribusi individual setiap teknik optimasi. Tabel 10 dan Tabel 11 menunjukkan hasil perbandingan keempat model pada data uji.

Tabel 10. Perbandingan Akurasi Model RM1-RM4

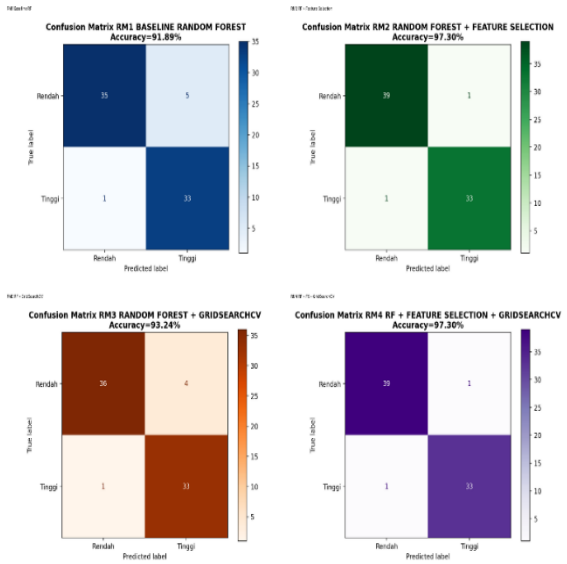
| Model | Konfigurasi | Fitur | Accuracy |
|---|---|---|---|
| RM1 | Baseline RF | 16 | 91,89% |
| RM2 | RF + Seleksi Fitur | 15 | 97,30% |
| RM3 | RF + GridSearchCV | 16 | 93,24% |
| RM4 | RF + FS + GSCV | 15 | 97,30% |

E-ISSN: 2580-1643

Tabel 11. Kappa, MCC, ROC-AUC, dan Gap Model

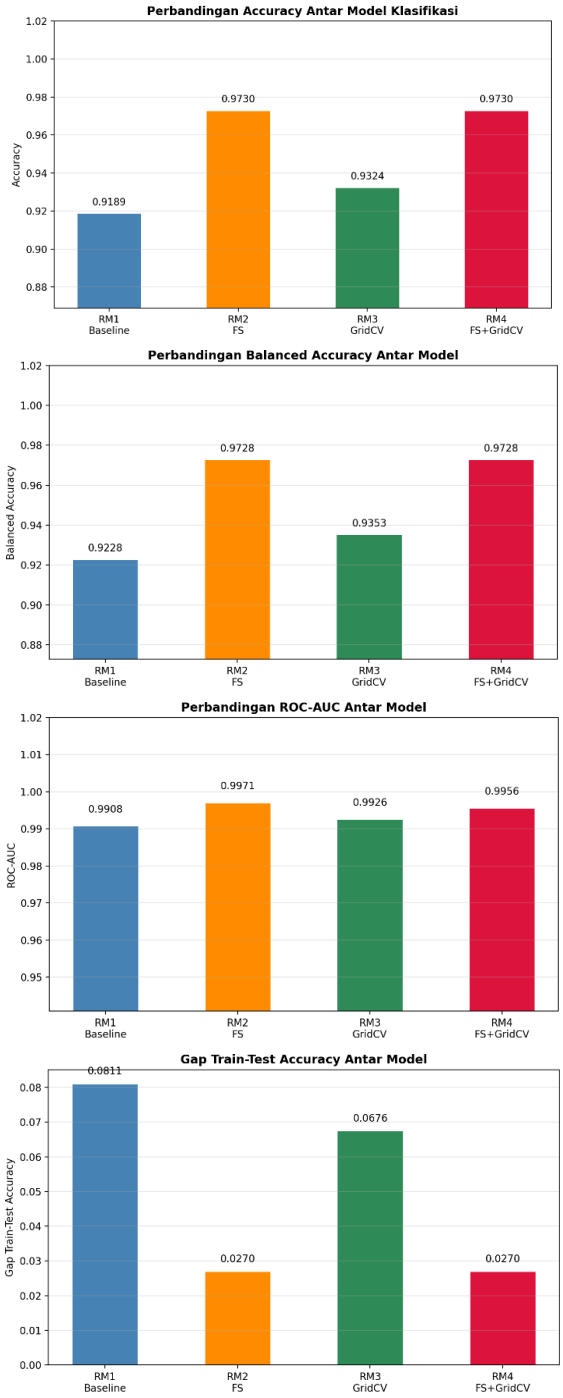
| Kappa | MCC | ROC-AUC | Gap |
|---|---|---|---|
| 0,8382 | 0,8431 | 0,9908 | 8,11% |
| 0,9456 | 0,9456 | 0,9971 | 2,70% |
| 0,8649 | 0,8677 | 0,9926 | 6,76% |
| 0,9456 | 0,9456 | 0,9956 | 2,70% |

RM2 dan RM4 menghasilkan accuracy tertinggi sebesar 97,30% dan balanced accuracy sebesar 97,28%, membuktikan bahwa model tidak hanya unggul secara keseluruhan tetapi juga seimbang dalam membedakan kelas Rendah dan Tinggi. Confusion matrix keempat model ditunjukkan pada Gambar 7.



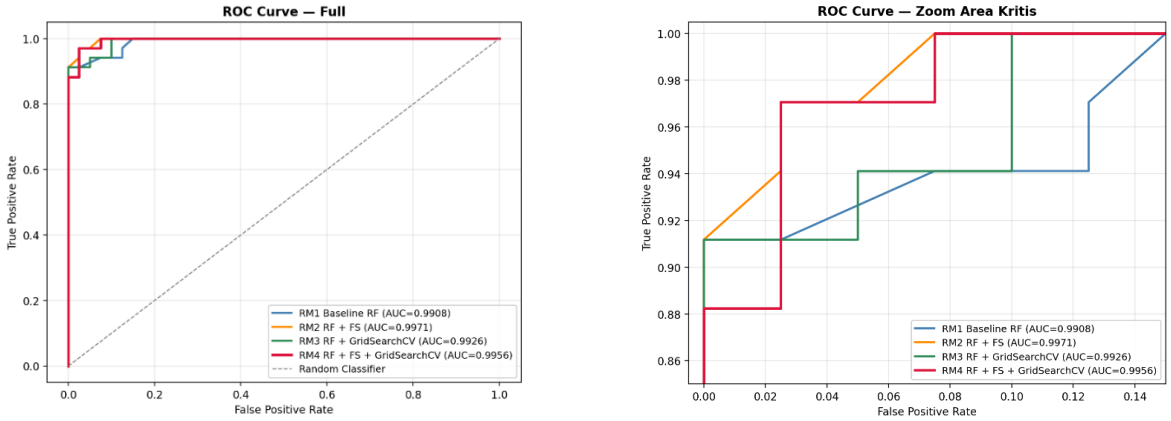

Gambar 7. Confusion Matrix Model RM1-RM4

Gap train-test RM2 dan RM4 sebesar 2,70% jauh lebih kecil dibandingkan RM1 (8,11%) dan RM3 (6,76%), menunjukkan bahwa seleksi fitur memberikan kontribusi terbesar terhadap pengurangan overfitting. *GridSearchCV* tanpa seleksi fitur (RM3) hanya meningkatkan accuracy dari 91,89% menjadi 93,24%, sedangkan seleksi fitur saja (RM2) langsung meningkat ke 97,30%. Temuan ini konsisten dengan (Gori & Hestiningtyas, 2024) bahwa seleksi fitur dan GridSearchCV memiliki efek saling melengkapi, serta sejalan dengan (Ristyawan *et al.*, 2025) bahwa preprocessing yang tepat lebih berpengaruh terhadap performa model dibandingkan hyperparameter tuning. Seluruh model memperoleh ROC-AUC di atas 0,99, mengkonfirmasi kemampuan pemisahan kelas yang sangat baik pada kedua kelas, sebagaimana ditunjukkan pada Gambar 8.




Gambar 8. Perbandingan Metrik Antar Model

E-ISSN: 2580-1643




Gambar 9. Kurva ROC Model Klasifikasi

Hasil evaluasi lebih lanjut menunjukkan bahwa model RM2 dan RM4 tidak hanya unggul dari sisi accuracy, tetapi juga konsisten pada tingkat per kelas. Tabel 12 dan Tabel 13 menyajikan nilai precision, recall, F1-score, dan support untuk masing-masing kelas pada kedua model tersebut.

Tabel 12. Precision dan Recall per Kelas

| Kelas | Precision | Recall |
|---|---|---|
| Rendah | 97,50% | 97,50% |
| Tinggi | 97,06% | 97,06% |
| Macro Average | 97,28% | 97,28% |
| Kelas | Precision | Recall |

Tabel 13. F1-Score dan Support per Kelas

| F1-Score | Support |
|---|---|
| 97,50% | 40 |
| 97,06% | 34 |
| 97,28% | 74 |

Pendekatan evaluasi multi-model ini terinspirasi dari (Marsya *et al.*, 2025) yang mengevaluasi RF secara komprehensif menggunakan confusion matrix dan classification report, serta (Ramadhan *et al.*, 2026) yang menggunakan macro F1-score sebagai metrik utama untuk menilai performa pada distribusi kelas yang tidak seimbang. Penggunaan MCC sebagai metrik tambahan juga didasarkan pada keunggulannya dalam menangani data tidak seimbang, dimana nilai MCC = 1 berarti prediksi sempurna dan 0 berarti setara tebakan acak (Ramadhan *et al.*, 2026). Gambar 9 menunjukkan kurva ROC seluruh model klasifikasi, yang mengkonfirmasi bahwa keempat model berada jauh di atas garis random classifier dengan AUC di atas 0,99.

### 3.2 Pembahasan

Hasil penelitian menunjukkan bahwa penggunaan model Random Forest dengan seleksi fitur berbasis feature importance dan optimasi hyperparameter GridSearchCV berhasil meningkatkan akurasi prediksi produksi padi di Indonesia. Model ini mampu mencapai nilai $R^2$ produksi sebesar 99,79% dan $R^2$ produktivitas sebesar 75,45%, yang menunjukkan bahwa meskipun ada dominasi luas panen aktual dalam konversi produksi, model ini tetap dapat memberikan estimasi yang akurat pada tingkat produktivitas. Hasil ini sejalan dengan penelitian oleh Manurung *et al.* (2025) yang juga menemukan bahwa model Random Forest dapat menangkap hubungan non-linier antara variabel iklim dan agronomis, serta menunjukkan keunggulan dibandingkan metode regresi linier. Selain itu, penelitian oleh Gori & Hestiningtyas (2024) menekankan pentingnya optimasi hyperparameter dalam meningkatkan performa model, yang juga tercermin dalam penelitian ini.

Penelitian ini berkontribusi pada literatur yang ada dengan mengintegrasikan seleksi fitur dan optimasi hyperparameter secara bersamaan, serta menggunakan data dari BPS dan NASA POWER yang lebih komprehensif dibandingkan dengan studi sebelumnya yang terbatas pada jumlah variabel iklim dan cakupan wilayah. Temuan ini memberikan wawasan berharga bagi pengambilan

E-ISSN: 2580-1643

keputusan di sektor pertanian untuk meningkatkan ketahanan pangan nasional melalui model yang lebih akurat dan dapat diandalkan.

## 4. Kesimpulan

Penelitian ini berhasil mengintegrasikan seleksi fitur berbasis feature importance internal Random Forest dengan optimasi hyperparameter GridSearchCV berskema TimeSeriesSplit untuk memvalidasi produksi padi pada 37 provinsi Indonesia periode 2018–2024. Proses penggabungan data BPS dan NASA POWER menghasilkan dataset akhir sebanyak 239 baris, yang dibagi secara kronologis menjadi 165 baris data latih (2018-2022) dan 74 baris data uji (2023-2024). Seleksi fitur mereduksi 32 fitur awal menjadi 16 fitur final, dengan Provinsi_Encoded, temp_range, daylight_mean, dan heat_index sebagai fitur paling berpengaruh, tanpa mengabaikan fitur agronomis wajib berupa curah hujan, kelembaban, dan suhu rata-rata. Optimasi GridSearchCV memperoleh konfigurasi terbaik dengan bootstrap=True, max_depth=None, max_features=0,5,min_impurity_decrease=0,0001, min_samples_leaf=1, min_samples_split=2, dan n_estimators=100, dengan Best CV MAE produktivitas sebesar 2,9347 ku/ha. Model akhir memperoleh $R^2$ produktivitas test sebesar 75,45% dan, setelah dikonversi menggunakan luas panen aktual, $R^2$ produksi test sebesar 99,79% dengan Adjusted $R^2$ sebesar 99,73%, RMSE sebesar 115.600,28 ton, MAE sebesar 61.194,75 ton, MAPE sebesar 6,95%, dan SMAPE sebesar 6,86%, jauh melampaui baseline Linear Regression. Hasil ini menunjukkan bahwa kombinasi Random Forest, seleksi fitur berbasis feature importance, dan optimasi GridSearchCV mampu menghasilkan model validasi produksi padi yang akurat.

## 5. Daftar Pustaka

Adin Musababa, M. (2024). Implementasi algoritma linear regression untuk prediksi produksi tanaman padi di Kabupaten Grobogan. *Data Sciences Indonesia (DSI)*, 3(2), 68–78. https://doi.org/10.47709/dsi.v3i2.3118.

Faizal, R., Abdullah, A., & Pangestika, M. W. (2025). Perbandingan Random Forest regressor dan decision tree regressor untuk prediksi hasil panen. *Jurnal CoSciTech (Computer Science and Information Technology)*, 6(2), 247–253. https://doi.org/10.37859/coscitech.v6i2.9966.

Fitri, E., & Nugraha, S. N. (2024). Optimasi kinerja linear regression, Random Forest regression, dan multilayer perceptron pada prediksi hasil panen. *INTI Nusa Mandiri*, 18(2), 210–217. https://doi.org/10.33480/inti.v18i2.5269.

Gori, T., & Hestiningtyas, A. (2024). Optimasi pemilihan fitur untuk prediksi penyakit jantung menggunakan algoritma genetika dan Random Forest. *The Indonesian Journal of Computer Science*, 13(5), 8491–8502. https://doi.org/10.33022/ijcs.v13i5.4214.

Handayani, D. N., & Qutub, S. (2025). Penerapan Random Forest untuk prediksi dan analisis kemiskinan. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 4(2), 405–412. https://doi.org/10.31004/riggs.v4i2.512.

Hutahaean, J., Yusup, D., & Purwantoro, P. (2024). Perbandingan metode linear regression, Random Forest, & k-nearest neighbor untuk prediksi produksi hasil panen padi di Provinsi Jawa Barat. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(3), 3895–3900. https://doi.org/10.36040/jati.v8i3.9821.

Jati, A. K., Utomo, B. R., Asmara, N. H. J., & Kusumastusi, R. (2025). Prediksi hasil panen untuk pertanian menggunakan model regresi machine learning. *Prosiding Seminar Nasional Amikom Surakarta*, 3, 186–195.

Kharisma, S. M. I., Hadiana, A. I., & Ramadhan, E. (2025). Model prediksi produksi padi berdasarkan curah hujan dan suhu menggunakan regresi linier berganda. *Jurnal Algoritma*, 22(2), 704–7014. https://doi.org/10.33364/algoritma/v.22-2.2793.

Maisat Eka Darmawan, Z., & Fauzan Dianta, A. (2023). Implementasi optimasi hyperparameter GridSearchCV pada sistem prediksi serangan jantung menggunakan SVM. *Teknologi: Jurnal Ilmiah Sistem Informas*, 13(1), 8–15. https://doi.org/10.26594/teknologi.v13i1.3098.

Manurung, D., Zealtiel, B., & Lubis, A. H. (2025). Prediksi produksi tanaman padi di Indonesia dengan menggunakan algoritma Random Forest regressor. *Journal of*

E-ISSN: 2580-1643

*Computing and Informatics Research*, 4(3), 337–345. https://doi.org/10.47065/comforch.v4i3.2125.

Marsya, N. D., Munadhil, M. M., Dzakyananta, M. A., Amaliah, K., & Rofianto, D. (2025). Penerapan algoritma Random Forest dalam prediksi emosi musik berdasarkan karakteristik fitur audio Spotify. *Jurnal Sains Informatika Terapan*, 4(2), 435–440. https://doi.org/10.62357/jsit.v4i2.648.

Masdian, A. R., Bashit, N., & Hadi, F. (2023). Analisis produktivitas padi menggunakan algoritma machine learning Random Forest di Kabupaten Batang tahun 2018—2022. *Elipsoida: Jurnal Geodesi dan Geomatika*, 6(1), 43–51. https://doi.org/10.14710/elipsoida.2023.19023.

Muchtar, I. R., & Afiyati, A. (2024). Comparison of linear regression and Random Forest algorithms for premium rice price prediction (Case Study: West Java). *Jurnal Indonesia Sosial Teknologi*, 5(7), 3122–3132. https://doi.org/10.59141/jist.v5i7.1184.

Nur Fauzi, N. P., Khomsah, S., & Putra Wicaksono, A. D. (2025). Penerapan feature engineering dan hyperparameter tuning untuk meningkatkan akurasi model Random Forest pada klasifikasi risiko kredit. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 12(2), 251–262. https://doi.org/10.25126/jtiik.2025128472.

Ramadhan, F., Herlambang, D., & Dipta, A. P. (2026). Prediksi status kesehatan berdasarkan gaya hidup menggunakan metode decision tree dan feature importance. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 4(4), 9616–9623. https://doi.org/10.31004/riggs.v4i4.5246.

Risanti, R., & Suhendar, H. (2024). Analisis model prediksi cuaca menggunakan support vector machine, gradient boosting, Random Forest, dan decision tree. *Prosiding Seminar Nasional Fisika*, 12, 119–127. https://doi.org/10.21009/03.1201.FA18.

Ristyawan, A., Nugroho, A., & Amarya, T. K. (2025). Optimasi preprocessing model Random Forest untuk prediksi stroke. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 12(1), 29–44. https://doi.org/10.35957/jatisi.v12i1.9587

Rizquina, A. Z., & Ratnasari, C. I. (2023). Implementasi web scraping untuk pengambilan data pada website e-commerce. *Jurnal Teknologi dan Sistem Informasi Bisnis*, 5(4), 377–383. https://doi.org/10.47233/jteksis.v5i4.913.

Selayanti, N., Putri, S. A., Kristanaya, M., Azzahra, M. P., Navsih, M. G., & Hindrayani, K. M. (2024). Penerapan machine learning algoritma Random Forest untuk prediksi penyakit jantung. *Prosiding Seminar Nasional Sains Data*, 4(1), 895–906. https://doi.org/10.33005/senada.v4i1.376.

Tantyoko, H., Sari, D. K., & Wijaya, A. R. (2023). Prediksi potensi gempa bumi Indonesia menggunakan metode Random Forest dan feature selection. *IDEALIS: InDonEsiA journaL Information System*, 6(2), 83–89. https://doi.org/10.36080/idealis.v6i2.3036

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.