**Journal of Applied Informatics and Computing (JAIC)**
Vol.9, No.6, December 2025, pp. 3479~3489
e-ISSN: 2548-6861


# Comparison of Multiple Linear Regression and Random Forest Methods for Predicting National Rice Production in Indonesia

**Sefrico Aji Nur Cahyo [1]*, Marcelinus Yosep Teguh Sulistyono [2]*,**
* Sistem Informasi, Universitas Dian Nuswantoro
112202206949@mhs.dinus.ac.id [1], teguh.sulistyono@dsn.dinus.ac.id [2],

**Article Info**

***Article history:***

Received 2025-09-30
Revised 2025-11-16
Accepted 2025-11-22

***Keyword:***

*Rice Production,*
*Multiple Linear Regression,*
*Prediction,*
*MAE,*
*RMSE,*
*AIC.*

**ABSTRACT**

Rice is a strategic commodity that plays an important role in maintaining national food security. However, rice production in Indonesia still fluctuates due to variations in harvest area, productivity, climate conditions, and differences in regional characteristics. This condition demands a predictive model capable of providing more accurate production estimates to support food policy planning. This research aims to predict national rice production by comparing two methods: Multiple Linear Regression and Random Forest Regression, using data from the Central Bureau of Statistics (BPS) and Nasa Power for the period 2018–2024. The analysis stages include data preprocessing, data exploration, categorical variable transformation, splitting data into training and testing sets, model training, and evaluation using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination ($R^2$). The research results show that harvested area is the most dominant factor influencing rice production, followed by productivity, year, and province. Based on the evaluation results, Random Forest provided the best performance with an MAE value of 40,599.94, an RMSE of 77,153.07, and an $R^2$ of 0.9991. The low error value and the proximity of the prediction to the actual data indicate that this model is better at capturing non-linear patterns and inter-regional variations compared to Multiple Linear Regression. Overall, Random Forest can be an effective method for predicting national rice production and can be further developed in subsequent research by incorporating climate variables or other external factors.



## I. PENDAHULUAN

Sebagai negara dengan basis pertanian yang kuat, Indonesia menempatkan padi sebagai komoditas strategis yang berperan besar dalam menjaga stabilitas ketahanan pangan negara. Berdasarkan informasi dari Badan Pusat Statistik (BPS), faktor luas panen, produktivitas, kondisi iklim, dan perbedaan karakteristik geografis antar daerah adalah beberapa faktor yang memengaruhi produksi padi dari tahun ke tahun. Fenomena ini menunjukkan bahwa produksi padi ditentukan oleh kombinasi banyak variabel yang saling berhubungan, bukan hanya satu faktor. Oleh karena itu, upaya perkiraan yang lebih presisi dibutuhkan untuk membantu perencanaan kebutuhan pangan di masa mendatang. [1] [2][3][4]

Faktor-faktor seperti luas panen berperan terhadap langsung atas jumlah hasil panen padi, oleh karena itu semakin luas lahan yang ditanami maka semakin besar pula hasil panen yang dapat diperoleh[5][6]. Akan tetapi, luas panen tidak selalu stabil dari tahun ke tahun, sebab sebagian lahan pertanian beralih fungsi menjadi pemukiman atau kawasan industri, sementara sebagian lainnya terpengaruh oleh perubahan pola tanam petani[7][8]. Faktor produktivitas juga sangat menentukan karena menggambarkan jumlah hasil yang diperoleh dari setiap hektar lahan. Tingkat produktivitas ini biasanya dipengaruhi oleh penggunaan varietas unggul, ketersediaan pupuk dan sarana produksi lainnya, serta penerapan teknologi budidaya yang digunakan[6], [9]. Selain itu, perbedaan antar provinsi turut memengaruhi hasil produksi, sebab setiap daerah memiliki kondisi geografis,

tingkat kesuburan tanah, ketersediaan air irigasi, serta infrastruktur pertanian yang berbeda[10]. Tidak hanya faktor internal seperti luas panen dan produktivitas, faktor eksternal berupa kondisi iklim juga berperan penting dalam menentukan tingkat produksi padi. Curah hujan yang terlalu rendah dapat menyebabkan kekeringan, sedangkan curah hujan berlebih dapat menimbulkan banjir yang menghambat pertumbuhan tanaman. Temperatur udara juga memengaruhi proses fisiologis tanaman padi, seperti fotosintesis dan pembentukan bulir[11]. Oleh karena itu, variasi faktor-faktor tersebut baik internal maupun eksternal menyebabkan produksi padi nasional berfluktuasi dari tahun ke tahun dan menuntut pendekatan analisis yang lebih adaptif untuk memprediksi hasil produksi secara akurat.

Masalah penurunan produksi padi di Indonesia erat kaitannya dengan alih fungsi lahan pertanian, degradasi kualitas tanah, serta tekanan lingkungan yang semakin meningkat. Beberapa penelitian menunjukkan bahwa laju konversi lahan sawah memiliki hubungan yang sangat kuat dengan penurunan produksi padi, di mana setiap hektar lahan yang beralih fungsi berkontribusi signifikan terhadap berkurangnya hasil panen[7][8]. Studi lain menegaskan bahwa transformasi penggunaan lahan dari pertanian menjadi perumahan serta sektor industri tidak hanya mengurangi ketersediaan pangan lokal, tetapi juga meningkatkan ketergantungan impor serta menimbulkan degradasi lingkungan dan ketimpangan sosial[8][6]. Selain itu, hasil penelitian juga membuktikan bahwa luas lahan sawah berperan langsung dalam menjaga kapasitas produksi beras nasional, penurunan luasan sawah berpotensi meningkatkan risiko ketergantungan impor dan mengancam stabilitas pangan jangka panjang. Kompleksitas permasalahan tersebut menunjukkan bahwa produksi padi ditentukan oleh beragam faktor yang saling berkaitan satu sama lain, baik dari aspek lahan, lingkungan, maupun iklim. Hubungan antarvariabel ini tidak selalu bersifat linier, sehingga pendekatan analisis konvensional sering kali kurang mampu menangkap pola perubahan yang dinamis dari waktu ke waktu[11][12]. Oleh karena itu, diperlukan model prediksi yang lebih adaptif dan mampu mempelajari hubungan non-linier antar faktor, seperti metode machine learning berbasis ensemble salah satunya adalah Random Forest untuk dibandingkan dengan model linier tradisional seperti Regresi Linier Berganda dalam memprediksi produksi padi nasional[13][12][14].

Random Forest merupakan metode pembelajaran mesin berbasis *ensemble* yang mampu menangkap hubungan kompleks antar variabel tanpa memerlukan asumsi linieritas. Dengan menggabungkan banyak pohon keputusan (*decision tree*), diharapkan metode tersebut mampu memproduksi prediksi yang lebih tepat dan stabil dibandingkan model regresi linier. Setiap tree dalam Random Forest dibangun memanfaatkan bagian-bagian data yang bervariasi, sehingga hasil akhirnya merupakan nilai tengah dari semua prediksi struktur pohon yang dihasilkan. Pendekatan ini membuat RF lebih adaptif terhadap variasi data dan mampu mengurangi risiko *overfitting* pada proses pemodelan[12][15]. Selain itu, penelitian ini juga menerapkan validasi silang (*cross-validation*) untuk memastikan keandalan model dalam menghadapi variasi data[16]. Teknik ini membagi dataset menjadi beberapa lipatan (*folds*) untuk dilatih dan diuji secara bergantian, sehingga evaluasi model menjadi lebih objektif dan tidak bergantung pada satu pembagian data tertentu. Kombinasi antara Regresi Linier Berganda dan Random Forest yang divalidasi melalui *cross-validation* diharapkan dapat memberikan hasil prediksi produksi padi nasional yang lebih akurat, stabil, dan representatif.

Sebagai upaya mengatasi permasalahan tersebut, salah satu alternatif solusi yang dapat diterapkan adalah dengan membangun model prediksi menggunakan pendekatan regresi dan pembelajaran mesin. Pada penelitian ini digunakan dua metode utama, yaitu Regresi Linier Berganda dan Random Forest Regression. Metode Regresi Linier Berganda memungkinkan pengujian keterkaitan antara produksi padi sebagai variabel dependen terhadap beberapa variabel independen seperti luas panen, produktivitas, curah hujan, temperatur, tahun, serta provinsi[2][5]. Sementara itu, Random Forest digunakan sebagai pembanding karena mampu menangkap hubungan non-linier dan interaksi kompleks antar variabel yang tidak dapat dijelaskan secara optimal oleh model linier. Proses analisis dilakukan melalui tahapan yang sistematis, meliputi pembersihan dan normalisasi data, konversi variabel kategorikal menggunakan *label encoding*, eksplorasi data (*exploratory data analysis*), pembagian data menjadi data latih dan uji, serta pembangunan model prediksi untuk kedua metode. Selanjutnya, performa model dievaluasi menggunakan ukuran kesalahan seperti Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi ($R^2$). Evaluasi tambahan dilakukan dengan teknik *cross-validation* untuk memastikan keandalan dan kestabilan hasil prediksi. Melalui langkah-langkah tersebut, diharapkan dapat diperoleh metode yang paling akurat dalam memprediksi produksi padi nasional.

Meskipun Regresi Linier Berganda memiliki keunggulan dalam hal kemudahan interpretasi dan struktur analisis yang jelas, model ini bergantung pada asumsi bahwa hubungan antara variabel independen dan variabel dependen bersifat linier. Dalam konteks produksi padi, asumsi tersebut dapat menjadi keterbatasan karena hasil produksi tidak hanya dipengaruhi oleh luas panen dan produktivitas, tetapi juga oleh faktor eksternal yang bersifat dinamis dan kompleks, seperti curah hujan dan temperatur. Perubahan kondisi iklim dan variasi geografis antar wilayah berpotensi menimbulkan hubungan yang non-linier terhadap hasil produksi. Oleh karena itu, penelitian ini tidak hanya membangun model Regresi Linier Berganda, tetapi juga membandingkannya dengan pendekatan non-linier, yaitu Random Forest Regression, untuk menilai sejauh mana perbedaan performa kedua metode dalam memprediksi produksi padi nasional. Analisis komparatif ini bertujuan untuk menentukan metode yang paling akurat dan stabil dalam menggambarkan hubungan antar variabel, serta untuk menguji apakah model linier masih relevan ketika faktor iklim dan dinamika

lingkungan turut dipertimbangkan dalam model prediksi produksi padi di Indonesia.

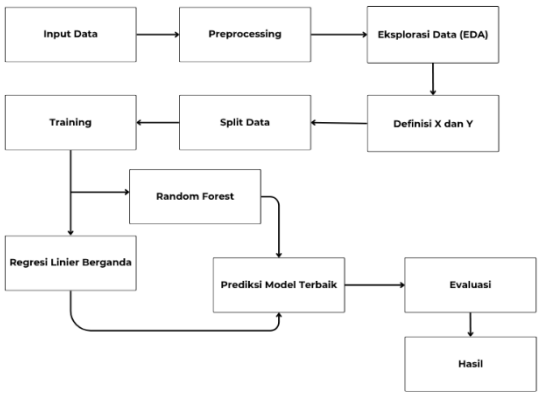
Penelitian ini bertujuan untuk mengolah dan mempersiapkan data produksi padi nasional dari Badan Pusat Statistik (BPS) agar dapat digunakan dalam pemodelan prediksi yang andal. Selanjutnya, penelitian ini membangun dua model prediksi, yaitu Regresi Linier Berganda dan Random Forest Regression, untuk memperkirakan produksi padi berdasarkan variabel luas panen, produktivitas, curah hujan, temperatur, tahun, dan provinsi. Kinerja kedua model dievaluasi dengan menggunakan metrik statistik seperti Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi ($R^2$), serta diuji menggunakan teknik *cross-validation* untuk memastikan keandalan hasil prediksi[12][17]. Hasil dari penelitian ini diharapkan dapat memberikan rekomendasi metode prediksi yang paling akurat dan stabil dalam menggambarkan hubungan antara faktor-faktor penentu produksi padi nasional[17][18]. Temuan penelitian ini juga dapat dimanfaatkan sebagai bahan pertimbangan dalam perencanaan produksi, kebijakan ketahanan pangan, serta pengambilan keputusan strategis untuk mendukung peningkatan produktivitas pertanian di Indonesia.

## II. METODE

Penelitian ini menggunakan pendekatan kuantitatif komparatif dengan dua metode utama, yaitu *Regresi Linier Berganda* dan *Random Forest Regression*. Kedua metode tersebut digunakan untuk memprediksi produksi padi nasional berdasarkan variabel luas panen, produktivitas, curah hujan, temperatur, tahun, dan provinsi[2][12]. Alur penelitian terdiri atas sembilan tahapan utama, yaitu input data, preprocessing untuk pembersihan dan normalisasi data, *exploratory data analysis* (EDA) untuk memahami pola dan korelasi antar variabel, penetapan variabel bebas (X) dan variabel terikat (Y), pembagian dataset dipisahkan menjadi bagian pelatihan dan bagian pengujian (*train-test split*), pelatihan model (*training*) menggunakan kedua metode, prediksi hasil produksi padi, evaluasi performa model menggunakan metrik MAE, RMSE, dan $R^2$, serta validasi hasil menggunakan *cross-validation* untuk menilai kestabilan dan keandalan model[12][16].

### A. Input Data

Data yang digunakan dalam penelitian ini merupakan data Gabah Kering Giling (GKG) dari setiap provinsi di Indonesia yang dihimpun dari Badan Pusat Statistik (BPS) serta NASA POWER (Prediction of Worldwide Energy Resources). Data dari BPS mencakup variabel provinsi, luas lahan panen padi (ha), jumlah hasil panen (ton) dan tingkat produktivitas (ha), serta tahun pengamatan, sedangkan data dari NASA POWER digunakan untuk memperoleh variabel iklim berupa curah hujan (mm) dan temperatur udara (°C)[3][19]. Periode data yang digunakan mencakup tahun 2018 hingga 2024, sehingga mampu merepresentasikan variasi kondisi produksi dan iklim antarwilayah di Indonesia[19]. Kombinasi antara data produksi padi dari BPS dan data iklim dari NASA POWER hal tersebut digunakan sebagai fondasi dalam penyusunan model prediksi menggunakan dua pendekatan, yaitu Regresi Linier Berganda dan Random Forest Regression, guna menganalisis pengaruh faktor-faktor internal (luas panen, produktivitas, tahun, dan provinsi) serta faktor eksternal (curah hujan dan temperatur) terhadap produksi padi nasional.




Gambar 1. Alur Penelitian

### B. Preprocessing

Sebelum dilakukan pemodelan dengan Regresi Linier Berganda dan Random Forest Regression, data terlebih dahulu diproses agar bersih dan seimbang. Dataset ini terdiri dari 170 observasi yang mencakup 34 provinsi dengan tujuh variabel utama, yaitu Provinsi, Luas Panen (ha), Produktivitas (ku/ha), Produksi (ton), Curah Hujan (mm), Temperatur (°C), dan Tahun[12]. Data diperoleh dari BPS dan NASA POWER dengan periode utama 2018 – 2024, namun untuk wilayah tertentu seperti Papua Papua pegunungn, Papua Barat Daya, Papua Selatan, dan Papua Tengah ketersediaan data berbeda pada rentang tahun 2022 – 2024. Untuk mengurangi ketidakseimbangan ini, dilakukan pemeriksaan distribusi data antar provinsi serta normalisasi numerik menggunakan StandardScaler[10][12]. Variabel kategorikal *Provinsi* diubah menjadi numerik melalui Label Encoding[7][8]. Semua proses dilakukan menggunakan Python agar data siap digunakan dalam tahap analisis dan pemodelan. Tahapan preprocessing yang dilakukan pada penelitian ini adalah sebagai berikut:

- Penanganan data Hilang (Missing Values)
  Baris data yang memiliki nilai kosong (missing) dihapus atau dilakukan imputasi (pengisian nilai) agar hasil analisis tidak menyimpang. Hal ini penting karena data yang tidak lengkap dapat memengaruhi akurasi model[12].
- Standarisasi Format Data
  Semua variabel numerik, seperti Produksi (ton) dan Luas Panen (ha), dipastikan berada dalam format

angka[12]. Tujuannya agar data dapat dihitung secara matematis oleh model regresi.

- Pengolahan Variabel Kategorikal
  Karena variabel Provinsi masih berupa data kategorikal, maka variabel itu perlu disesuaikan ke format numerik. Transformasi dilakukan dengan teknik *One-Hot Encoding*, di mana setiap kategori wilayah direpresentasikan dalam variabel biner[7][8][6].
  Persamaan konversi variabel kategorikal dengan One-Hot Encoding dalam model linear regresi berganda dapat dituliskan sebagai berikut[12][15]:

$$\gamma = a + b_1\mathrm{X}_1 + b_2\mathrm{X}_2 + \cdots + b_n\mathrm{X}_n \quad [16] \qquad (1)$$

Dimana

$\gamma$ = Variabel tidak bebas (nilai yang diprediksikan)
X = Variabel bebas
$a$ = Konstanta
$b$ = Koefisie

Jika variabel memiliki satuan berbeda dan skala yang jauh, normalisasi menggunakan Standard Scaler bisa dilakukan agar model lebih stabil saat training[17] [11].

$$x'_i = \frac{x_i - \mu}{\sigma} \qquad [9], [10] \ (2)$$

Dimana

$x_i$ = nilai asli dari fitur
$\mu$ = rata-rata fitur (mean)
$\sigma$ = standar deviasi fitur

### C. Eksplorasi Data (EDA)

Sebelum model dibangun, lakukan eksplorasi data secara visual dan numerik: plot distribusi, scatter plot, dan matriks korelasi guna memahami pola dan hubungan antar fitur. EDA bukan hanya formalitas dari sini kamu bisa menemukan outlier, tren, atau hubungan linear yang mungkin tidak terlihat sekilas.

### D. Definisi X dan Y

Dalam penelitian ini, variabel yang digunakan dibagi menjadi dua kelompok, yaitu variabel target (*dependent variable*) dan variabel prediktor (*independent variables*). Variabel target merupakan variabel utama yang akan diprediksi, sedangkan variabel prediktor adalah variabel yang diduga memiliki pengaruh terhadap variabel target [2].

Berikut merupakan uraian definisi variabel yang dipakai dalam penelitian ini:

- Variabel Target (Y): Produksi padi (ton), yaitu jumlah total hasil produksi padi dalam satuan ton yang menjadi fokus utama prediksi[1] [2].
- Variabel Prediktor (X):
  1) Luas Panen (ha): Total luas lahan panen padi dalam satuan hektar.
  2) Produktivitas (ku/ha): Tingkat hasil produksi per hektar dalam satuan kuintal per hektar.
  3) Tahun: Periode waktu pengamatan data (time series).
  4) Provinsi (encoded): Wilayah administratif tempat produksi padi, yang dikonversi menjadi nilai numerik melalui proses encoding supaya dapat diolah oleh model regresi [1].
  5) Curah Hujan: ketinggian Tingkat curah hujan
  6) Temperatur: kelembabasn atau suhu udara 2 meter di atas tanah

### E. Split Data

Pada tahap ini merupakan pembagian dataset menjadi dua: 80 % untuk pelatihan (*training set*) dan 20 % untuk pengujian (*testing set*)[12]. Pemisahan ini perlu acak untuk menjaga representasi yang seimbang.

### F. Training

Dalam tahap ini, bangun model regresi linier berganda di data latih dengan meminimalkan error residual. Pastikan diperiksa juga asumsi regresi: linearitas, homoskedastisitas, distribusi error normal, dan tidak ada multikolinearitas yang mengganggu.

### G. Regresi Linier Berganda

Multiple linear regression merupakan merupakan pendekatan statistik yang diterapkan guna menganalisis keterkaitan hubungan antara satu variabel dependen dengan dua atau lebih variabel independen. Model ini berasumsi bahwa hubungan antar variabel bersifat linier, artinya setiap perubahan pada variabel independen akan menyebabkan perubahan proporsional pada variabel dependen. Dalam penelitian ini, Regresi Linier Berganda digunakan untuk memprediksi produksi padi (ton) berdasarkan variabel Luas Panen (ha), Produktivitas (ku/ha), Curah Hujan (mm), Temperatur (°C), Tahun, dan Provinsi[2].

Rumus:

$$y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (3)$$

$$b_1 = \frac{(\Sigma {X_2}^2)(\Sigma X_1 y) - (\Sigma X_2 y)(\Sigma X_2 X_2)}{(\Sigma {X_1}^2)(\Sigma {X_2}^2) - (\Sigma X_1 X_2)^2} \qquad (4)$$

$$b_2 = \frac{(\Sigma {X_2}^2)(\Sigma X_2 y) - (\Sigma X_1 y)(\Sigma X_1 X_2)}{(\Sigma {X_1}^2)(\Sigma {X_2}^2) - (\Sigma X_1 X_2)^2} \qquad (5)$$

$$a = \frac{\Sigma \mathrm{Y} - (b_1 \Sigma X_1)(b_2 \Sigma X_2)}{n} \qquad (6)$$

Dimana

$y$ = Variabel terikat (hasil yang diprediksikan)[10]
$X$ = Variabel independen
$a$ = Nilai Konstanta
$b$ = Koefisien regresi

### H. Random Forest

Random Forest merupakan metode ensemble learning yang menggabungkan banyak pohon keputusan (decision trees) untuk menghasilkan prediksi yang lebih akurat dan stabil. Setiap pohon dilatih menggunakan sampel data dan

fitur yang dipilih secara acak (bootstrap sampling), kemudian hasil prediksi akhir diperoleh dari rata-rata seluruh pohon. Pada studi ini, algoritma Random Forest dimanfaatkan untuk melakukan prediksi produksi padi (ton) berdasar pada Luas Panen (ha), Produktivitas (ku/ha), Curah Hujan (mm), Temperatur (°C), Tahun, dan Provinsi. Metode ini dapat mengidentifikasi hubungan non-linier dan interaksi kompleksitas hubungan antara variabel yang tidak dapat dijelaskan sepenuhnya oleh model linier, serta mengurangi risiko overfitting sehingga menghasilkan prediksi yang lebih andal[20].

Rumus umum dari model Random Forest dapat dituliskan sebagai berikut:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad [20](10)$$

Dimana

$\hat{y}$ = Prediksi akhir
$T$ = Jumlah pohon (trees)
$h_t(x)$ = Prediksi dari pohon ke-$t$
$\sum_{t=1}^{T}$= Penjumlahan semua prediksi
$\frac{1}{T}$ = Rata-rata aritmetika

*I. Evaluasi*

Tujuan evaluasi model adalah untuk menilai tingkat kedekatan antara output prediksi dengan nilai aktual pada data uji. Penilaian performa model dilakukan dengan memanfaatkan sejumlah ukuran statistik, seperti Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), serta nilai Koefisien Determinasi ($R^2$). Ketiga metrik tersebut berfungsi untuk melihat seberapa besar tingkat error dan tingkat ketepatan prediksi yang dihasilkan model semakin kecil nilai MAE dan RMSE serta semakin tinggi nilai $R^2$, maka semakin baik kemampuan model dalam memprediksi produksi padi. Selain itu, untuk memastikan keandalan beserta potensi penerapan secara luas model terhadap data baru, penelitian ini juga menerapkan validasi silang (cross-validation)[13]. Teknik ini membagi dataset menjadi beberapa bagian (fold) untuk dilatih dan diuji secara bergantian, sehingga evaluasi tidak bergantung pada satu kali pembagian data saja. Pendekatan ini membantu mengurangi risiko *overfitting* serta memberikan gambaran yang lebih objektif terhadap performa model, baik untuk Regresi Linier Berganda maupun Random Forest Regression. Evaluasi dilakukan dengan menghitung beberapa metrik, yaitu:[2][13]

- *Mean Absolute Error* (MAE):
  mengukur rata-rata selisih absolut antara nilai aktual ($y_i$) dan nilai prediksi ($\hat{y}_i$). MAE menunjukkan seberapa besar kesalahan rata-rata prediksi tanpa memperhatikan arah (positif atau negatif). [18]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

Dimana

$n$ = Jumlah data
$y_i$ = Nilai aktual ke- $i$
$\hat{y}_i$ = Nilai prediksi ke- $i$
$|y_i - \hat{y}_i|$ = Besarnya selisih absolut antara data aktual dan hasil prediksi

- *Root Mean Squared Error* (RMSE):
  merupakan nilai akar dari kuadrat berdasarkan MSE. RMSE berfungsi untuk mengukur error prediksi dihitung rata-rata dalam satuan yang identik dengan data asli, namun metrik ini lebih rentan dipengaruhi oleh nilai pencilan.[18][2]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

Dimana

$n$ = banyak data
$y_i$ = nilai aktual ke- $i$
$\hat{y}_i$ = nilai prediksi ke- $i$
$(y_i - \hat{y}_i)^2$ = selisih absolut antara nilai aktual dan prediksi

- Koefisien Determinasi ($R^2$)
  Koefisien determinasi ($R^2$)[13] adalah ukuran statistik yang menunjukkan seberapa baik model regresi dapat menjelaskan variasi data yang diamati. Dengan kata lain, $R^2$ menggambarkan tingkat pengaruh variabel bebas (X) terhadap variabel terikat (Y).

$$R^2 = 1\frac{SS_{res}}{SS_{tot}} \quad (9)$$

Dimana

$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ = *residual sum of squares* (galat model)
$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ = *total sum of squares* (variansi total)
$\hat{y}_i$ = nilai prediksi model
$\bar{y}_i$ = rata-rata observasi
$n$ = jumlah sampel

## III. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil dari penelitian yang dilakukan melalui berbagai tahapan analisis, mulai dari preprocessing data hingga evaluasi model regresi linier berganda. Hasil ini mencakup gambaran dataset, eksplorasi pola data, pemodelan regresi, dan hasil prediksi. Selanjutnya, diskusi dilakukan untuk menginterpretasikan hasil penelitian dan menghubungkannya dengan teori dan hasil penelitian sebelumnya. Tujuan dari diskusi ini adalah untuk mendapatkan pemahaman yang lebih komprehensif tentang komponen yang memengaruhi produksi padi nasional.

*A. Dataset*

Dataset ini merupakan data Gabah Kering Giling (GKG) yang dikumpulkan oleh Badan Pusat Statistik (BPS) dalam rentang tahun 2018 hingga 2024. Dataset mencakup beberapa variabel utama, yaitu provinsi, luas panen, produktivitas, produksi, dan tahun. Data ini diperoleh melalui laporan Statistik Pertanian yang dikirimkan oleh Kepala Cabang

Dinas (KCD) di tingkat kecamatan, sehingga memiliki tingkat akurasi yang baik karena bersumber langsung dari satuan kerja daerah.Untuk keperluan analisis dan penulisan dalam jurnal ilmiah, nama-nama atribut pada dataset ini kemudian disingkat agar lebih ringkas, konsisten, dan mudah digunakan dalam model statistik maupun visualisasi data. Adapun penyingkatan atribut tersebut adalah sebagai berikut Provinsi (Prov), Luas Panen (LP), Produktivitas (Prodktv), Produksi (Prod), Curah Hujan (CH), Temperatur (Temp), Tahun (Thn).

TABEL I
DATA GABAH KERING(2018 - 2024)

| Prov | LP (ha) | Prodktv(ha) | Prod (Ton) | CH | Temp | Thn |
|---|---|---|---|---|---|---|
| Aceh | 329.1516 | 56 | 1.861.567 | 9 | 21 | 2018 |
| Sumatra Utara | 407.176 | 52 | 2.108.284 | 12 | 23 | 2018 |
| Sumatra Barat | 313.051 | 47 | 1.483.076 | 10 | 24 | 2018 |
| … | … | … | … | … | … | … |
| Papua Selatan | 47.169 | 46 | 21.780 | 12 | 26 | 2024 |
| Papua Tengah | 1.436 | 42 | 6.072 | 12 | 20 | 2024 |
| Papua Pegunungan | 10 | 44 | 42 | 5 | 18 | 2024 |

## B. *Preprocessing*

Tahap preprocessing dilakukan untuk menyiapkan data agar layak digunakan dalam pemodelan. Dataset yang digunakan terdiri dari 170 observasi dari 34 provinsi di Indonesia dengan tujuh variabel utama: Provinsi, Luas Panen, Produktivitas, Produksi, Curah Hujan, Temperatur, dan Tahun. Mengingat beberapa wilayah di bagian timur, seperti Papua, hanya memiliki ketersediaan data pada tahun 2022–2024, dilakukan pemeriksaan distribusi data untuk mengidentifikasi potensi ketidakseimbangan antarwilayah. Selanjutnya, data dibersihkan dengan menghapus nilai hilang dan baris duplikat, serta memastikan seluruh variabel numerik berada dalam format angka dengan menghilangkan karakter yang tidak relevan, seperti spasi, koma, dan tanda strip.

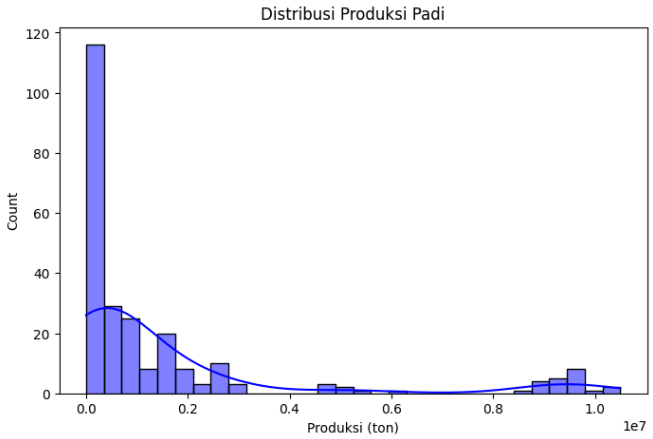
Setelah data dinyatakan valid, variabel kategorikal Provinsi dikonversi menjadi nilai numerik menggunakan Label Encoding agar dapat diproses oleh algoritma machine learning. Variabel numerik prediktor seperti Luas Panen, Produktivitas, Curah Hujan, dan Temperatur kemudian dinormalisasi menggunakan StandardScaler untuk memastikan semua variabel berada dalam skala yang sebanding dan tidak menimbulkan dominasi fitur akibat perbedaan satuan. Sementara itu, variabel Produksi (ton) tetap dibiarkan dalam bentuk aslinya karena berperan sebagai target model. Dengan langkah preprocessing ini, dataset menjadi lebih bersih, konsisten, dan siap digunakan dalam tahap pemodelan Regresi Linier Berganda dan Random Forest Regression.

TABEL II
HASIL PREPROCESSING

| Prov_enc | LP (ha) | Prodktv(ha) | Prod (Ton) | CH | Temp | Thn |
|---|---|---|---|---|---|---|
| 0 | -0.142 | 1.156 | 1861567 | 0.622 | -1.942 | 2018 |
| 38 | -0.096 | 0.682 | 2108285 | 2.055 | -1.281 | 2018 |
| 36 | -0.305 | 0.261 | 1483077 | 1.098 | -0.739 | 2018 |
| … | … | … | … | … | … | … |
| 28 | -0.305 | 0.142 | 217790 | 1.999 | 0.458 | 2024 |
| 29 | -0. 331 | -0. 240 | 6072 | 1.935 | -2.701 | 2024 |
| 27 | -0. 332 | -0. 084 | 42 | -0.905 | -3.400 | 2024 |

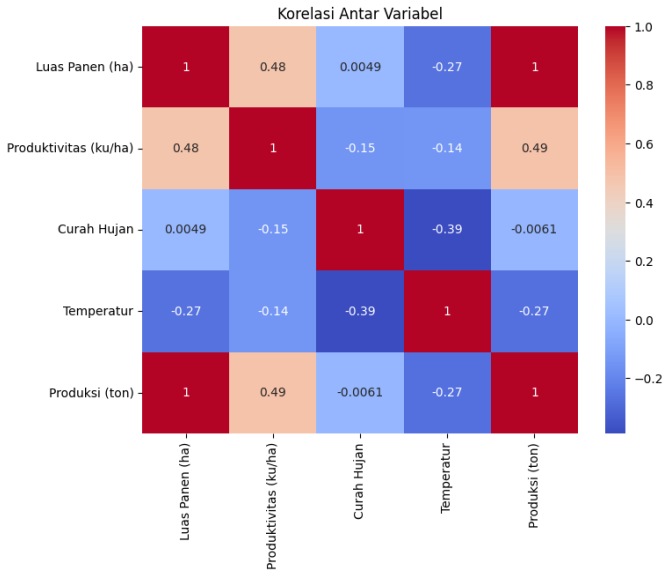
## C. *Exploratory Data Analysis EDA*

Exploratory Data Analysis (EDA) dilakukan setelah data melalui tahap preprocessing untuk mempelajari pola distribusi data dan hubungan antar variabel. Analisis ini bertujuan untuk menyajikan deskripsi awal mengenai karakteristik data sebelum dilakukan proses pemodelan.




Gambar 2. Distribusi Produk padi

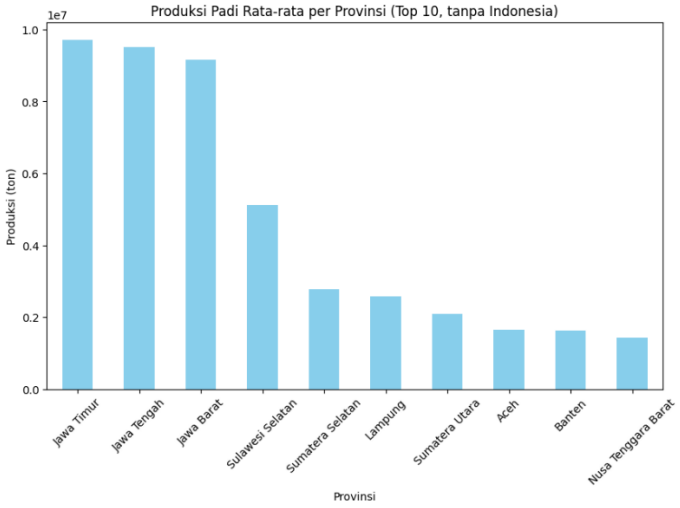
Histogram distribusi produksi padi di Indonesia ditunjukkan pada Gambar 3. Hasil menunjukkan bahwa distribusi data condong ke kanan (skewed ke kanan). Hal tersebut menggambarkan bahwa tidak banyak provinsi yang mencatatkan tingkat produksi yang tinggi, yang berdampak pada distribusi, sedangkan sebagian besar provinsi memiliki tingkat produksi yang relatif kecil hingga sedang. Pola ini menunjukkan bahwa kontribusi total produksi padi negara bervariasi antar provinsi.




Gambar 3.Matrik Korelasi Antara Variabel

Hasil analisis korelasi pada Gambar 3 menunjukkan hubungan antar variabel Luas Panen, Produktivitas, Curah Hujan, Temperatur, dan Produksi. Nilai korelasi yang ditampilkan dalam heatmap mengindikasikan bahwa Luas Panen (ha) memiliki hubungan paling kuat dengan Produksi (ton), ditunjukkan oleh nilai korelasi mendekati 1, sehingga produksi yang dihasilkan cenderung meningkat seiring bertambahnya luas lahan panen. Korelasi positif juga terlihat antara Produktivitas (ku/ha) dan Produksi (ton) dengan nilai sekitar 0.49, yang berarti peningkatan produktivitas turut mendorong peningkatan produksi, meskipun besarnya pengaruh tidak sekuat luas panen.

Sementara itu, variabel iklim menunjukkan hubungan yang relatif lemah terhadap produksi. Curah Hujan memiliki korelasi sangat kecil dengan Produksi (≈ –0.006), sehingga perubahan curah hujan dalam dataset ini tidak menunjukkan hubungan langsung terhadap variasi produksi. Temperatur memiliki korelasi negatif dengan Produksi (≈ –0.27), yang mengindikasikan bahwa kenaikan suhu cenderung menurunkan hasil produksi, meskipun tingkat pengaruhnya tergolong lemah. Hubungan antara variabel iklim juga terlihat, di mana Curah Hujan dan Temperatur memiliki korelasi negatif sedang (≈ –0.39), mencerminkan pola iklim di mana wilayah dengan curah hujan tinggi cenderung memiliki temperatur lebih rendah. Secara keseluruhan, heatmap memperlihatkan bahwa faktor internal seperti luas panen dan tingkat produktivitas memberikan dampak yang lebih dominan pada hasil panen padi dibanding pengaruh eksternal seperti curah hujan dan temperatu




Gambar 4.Produksi Padi Rata-rata per Provinsi

Gambar 4. menunjukkan hasil analisis, yang mengambil sepuluh provinsi teratas berdasarkan jumlah produksi padi, dan menampilkan bar chart yang menunjukkan produksi padi rata-rata per provinsi. Jumlah ini tidak termasuk agregat nasional Indonesia. Grafik ini menunjukkan perbedaan kontribusi produksi antarprovinsi; beberapa provinsi menunjukkan tingkat produksi yang lebih tinggi daripada yang lain.

Secara umum, grafik menunjukkan bahwa provinsi dengan lahan pertanian yang luas dan produktivitas tinggi menempati posisi teratas. Ini menunjukkan bahwa produksi padi terkonsentrasi di daerah tertentu, yang berfungsi sebagai pusat utama untuk memenuhi kebutuhan beras nasional. Nilai produksi yang dicantumkan di atas setiap batang memudahkan melihat besaran produksi rata-rata tiap provinsi, yang membuat pemahaman pola distribusi produksi lebih mudah.

Dari hasil visualisasi ini dapat dilihat bahwa meskipun produksi padi tersebar di berbagai provinsi, terdapat beberapa daerah yang mendominasi produksi nasional. Kondisi ini penting untuk diperhatikan dalam perumusan kebijakan pangan, karena ketergantungan pada provinsi tertentu berpotensi menimbulkan risiko apabila terjadi penurunan produksi akibat factor cuaca, hama, atau bencana alam di wilayah tersebut.

### *D. Definisi X dan Y*

Setelah Setelah proses eksplorasi data dilakukan, tahap berikutnya adalah menetapkan variabel yang digunakan dalam pemodelan. Penelitian ini mengelompokkan variabel menjadi dua jenis, yaitu variabel independen (X) dan variabel dependen (Y). Variabel independen mencakup Luas Panen (ha) sebagai indikator luas lahan yang dipanen di setiap provinsi, Produktivitas (ku/ha) sebagai ukuran hasil yang diperoleh per hektar, Tahun untuk menangkap perubahan produksi dari waktu ke waktu, serta Provinsi_enc, yaitu

representasi numerik dari variabel provinsi yang diolah menggunakan teknik *Label Encoding*. Transformasi ini diperlukan agar data kategorikal dapat diproses oleh algoritma regresi.

Variabel dependen dalam penelitian ini adalah Produksi (ton), yakni total hasil produksi padi yang menjadi fokus prediksi. Dengan susunan variabel tersebut, Penggunaan model Regresi Linier Berganda dianggap relevan karena mampu menjelaskan korelasi linear di antara berbagai variabel independen dengan produksi padi. Namun, untuk menangkap pola yang bersifat non-linier dan interaksi kompleks antar variabel, penelitian ini juga menggunakan Random Forest Regression sebagai metode pembanding. Penetapan variabel ini selaras dengan tujuan penelitian, yaitu mengidentifikasi faktor-faktor yang memengaruhi produksi padi serta menyusun model prediksi berbasis data historis dari berbagai provinsi dan tahun pengamatan.

### *E. Split Data*

Setelah menentukan variabel independen (X) dan dependen (Y), langkah berikutnya adalah membagi dataset menjadi set latihan (training set) dan set uji (testing set). proses pemisahan data dilakukan menggunakan fungsi *train_test_split* dari library scikit-learn. Pada studi ini, data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Untuk menjaga konsistensi serta memungkinkan hasil yang dapat direproduksi, digunakan pengaturan *random_state* dengan nilai 42.. Pelatihan model dilakukan dengan memanfaatkan data latih (X_train, y_train). Ini membangun hubungan matematis antara variabel independen (Luas Panen, Produktivitas, Tahun, Curah Hujan, Temperatur dan Provinsi_enc) dan variabel dependen (Produksi). Secara bersamaan, data uji (X_test, y_test) dimanfaatkan untuk menilai kinerja model pada data yang belum pernah digunakan saat pelatihan. Proses pembagian ini sangat penting agar model tidak sekadar mengingat pola data dari data latih (overfitting), tetapi juga mampu melakukan generalisasi ketika dihadapkan pada data baru, sehingga prediksi yang dihasilkan lebih akurat dan dapat diandalkan.

### *F. Training*

Proses pelatihan model dilakukan dengan memanfaatkan data latih yang telah dipisahkan pada tahap sebelumnya. Pada metode Regresi Linier Berganda, model dibangun dengan mempelajari hubungan linear antara variabel predictor Luas Panen, Produktivitas, Tahun, Curah Hujan, Temperatur, dan Provinsi_enc terhadap variabel target, yaitu Produksi padi. Pendekatan ini memungkinkan analisis kontribusi masing-masing variabel secara simultan terhadap perubahan nilai produksi. Selain itu, metode Random Forest Regression juga dilatih menggunakan data latih yang sama. Model ini bekerja dengan membangun banyak pohon keputusan secara acak (ensemble) sehingga mampu menangkap pola non-linier dan interaksi kompleks antar variabel. Kedua model tersebut dilatih untuk mengenali pola historis yang terkandung di dalam data, sementara data uji digunakan untuk menilai kemampuan model dalam menghasilkan prediksi pada data yang tidak digunakan selama proses pelatihan. Dengan demikian, proses training tidak hanya membangun model, tetapi juga memastikan bahwa model memiliki kemampuan generalisasi yang baik.

### *G. Pediksi*

Sesudah kedua model Regresi Linier Berganda maupun Random Forest disusun menggunakan data pelatihan, langkah selanjutnya adalah menghasilkan prediksi pada data pengujian untuk mengevaluasi sejauh mana masing-masing model mampu memperkirakan nilai produksi padi. Proses ini penting untuk melihat bagaimana model merespons pola variabel independen seperti luas panen, produktivitas, tahun, serta wilayah provinsi ketika diuji menggunakan data yang tidak termasuk dalam proses pelatihan model. Nilai prediksi kemudian dibandingkan dengan data aktual produksi padi guna mengetahui tingkat kedekatan antara hasil estimasi model dan kondisi sebenarnya.

Hasil pengujian menunjukkan bahwa kedua model mampu menghasilkan nilai prediksi yang cukup mendekati data aktual, namun performanya tidak sama. Regresi Linier Berganda memperlihatkan pola prediksi yang mengikuti tren umum tetapi kurang akurat pada daerah dengan variasi data yang tinggi atau hubungan yang tidak bersifat linier penuh. Sebaliknya, Random Forest menghasilkan prediksi dengan tingkat ketepatan dan kestabilan yang lebih tinggi. Keunggulan ini muncul karena Random Forest mampu menangkap pola non-linier, interaksi antarvariabel, serta perbedaan karakteristik antarprovinsi yang tidak dapat ditangani secara optimal oleh model linier. Dengan demikian, berdasarkan perbandingan nilai error dan kedekatan prediksi dengan data aktual, Random Forest dinilai sebagai metode yang lebih baik dalam memodelkan dan memprediksi produksi padi pada penelitian ini. Dengan mempertimbangkan tingkat galat yang lebih kecil dan kedekatan prediksi yang lebih baik terhadap data aktual, Random Forest menjadi model yang paling unggul dalam penelitian tersebut. Sehingga, pada bagian selanjutnya ditampilkan hasil perbandingan performa serta visualisasi prediksi dari model terbaik tersebut.

TABEL III
HASIL PREDIKSI PRODUKSI PADI

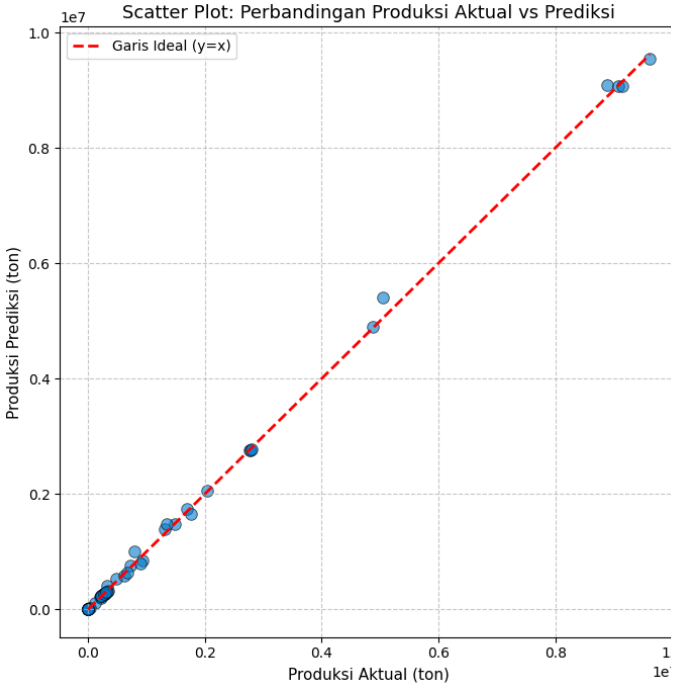
| No | Produksi Prediksi(ton) | Produksi Prediksi(ton) |
|---|---|---|
| 1 | 281,610.09 | 271,724.277 |
| 2 | 288,810.52 | 267,185.316 |
| 3 | 345,050.37 | 314,915.931 |
| 4 | 5,054,166.96 | 5,408,258.356 |
| 5 | 3,249.47 | 2,612.288 |

Tabel tersebut menampilkan perbandingan antara nilai produksi padi aktual dan hasil prediksi model Random Forest

untuk beberapa sampel data. Secara umum, nilai prediksi berada cukup dekat dengan nilai aktual, baik pada data dengan skala kecil maupun besar. Misalnya, pada sampel pertama dan kedua, perbedaan antara nilai asli dan hasil prediksi cukup minim, menunjukkan bahwa model dapat merepresentasikan pola data secara akurat. Pada sampel dengan nilai produksi yang sangat tinggi atau sangat rendah, model juga tetap memberikan estimasi yang searah dengan data asli. Tabel ini menunjukkan contoh hasil prediksi dari model terbaik, yaitu Random Forest, setelah dibandingkan dengan Regresi Linier Berganda. Nilai prediksi yang ditampilkan terlihat cukup dekat dengan data aktual, yang menegaskan keunggulan model ini dalam menghasilkan estimasi yang lebih presisi dibandingkan metode lainnya.

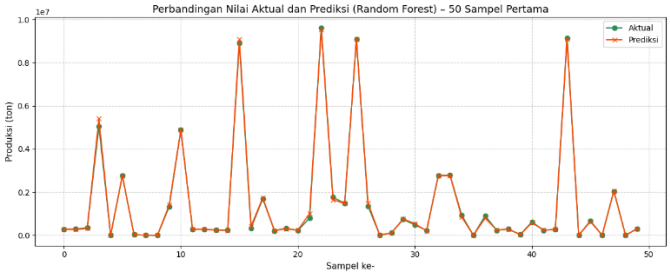
Visualisasi Hasil Prediksi

- Scatter Plot (Aktual vs Prediksi)
  Nilai produksi aktual dan hasil prediksi model berhubungan, seperti yang ditunjukkan oleh plot dispersi pada Gambar 5. Mayoritas titik data terletak di sekitar garis diagonal merah (y = x), yang menandakan bahwa model mampu memberikan prediksi yang cukup mendekati nilai aktual.




Gambar 5.Scatter Plot Perbandingan Produksi Aktual dan Prediksi

- Line Plot (Perbandingan Antara Data Aktual dan Hasil Prediksi)
  Perbandingan nilai aktual dengan hasil prediksi pada lima puluh sampel pertama ditampilkan dalam bentuk plot garis pada Gambar 6. Walaupun ada beberapa titik yang sangat berbeda, pola garis prediksi umumnya mengikuti tren garis aktual. Kondisi ini menunjukkan bahwa model menangkap pola data dengan baik, tetapi belum sepenuhnya sempurna untuk menunjukkan perbedaan produksi padi di setiap sampel.




Gambar 6. Line Plot Perbandingan Produksi Aktual dan Prediksi

*H. Evaluasi*

Setelah model Regresi Linier Berganda dan Random Forest Regression dilatih, langkah berikutnya adalah mengevaluasi kemampuan keduanya dalam memprediksi produksi padi. Evaluasi dilakukan menggunakan tiga metrik utama, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi ($R^2$). MAE mengukur rata-rata selisih absolut antara nilai aktual dan prediksi, sedangkan RMSE memberikan bobot lebih besar pada kesalahan yang ekstrem. Kedua metrik ini menunjukkan seberapa besar error prediksi yang dihasilkan model. Nilai $R^2$ dipakai untuk mengukur sejauh mana model dapat menggambarkan keragaman data yang ada semakin mendekati 1, semakin baik kinerja model.

Melalui ketiga indikator tersebut, performa kedua model dapat dibandingkan secara objektif. Regresi Linier Berganda bekerja berdasarkan pola hubungan linear antarvariabel, sementara Random Forest mampu menangkap pola non-linear melalui kombinasi banyak pohon keputusan. Oleh karena itu, perbedaan nilai MAE, RMSE, dan $R^2$ dari kedua model mencerminkan kemampuan masing-masing metode dalam memahami kompleksitas dan variasi data produksi padi nasional.

TABEL IV
HASIL EVALUASI

| Linear Regression | |
|---|---|
| **MAE** | 129878.89 |
| **RMSE** | 184311.20 |
| **R2** | 0.9947 |
| Random Forest Regression | |
| **MAE** | 40599.94 |
| **RMSE** | 77153.07 |
| **R2** | 0.9991 |

Tabel tersebut memperlihatkan perbandingan kinerja antara Regresi Linier Berganda dan Random Forest Regression berdasarkan tiga metrik evaluasi, yaitu MAE, RMSE, dan $R^2$. Secara teknis, nilai MAE dan RMSE pada model linear masih cukup tinggi sehingga menunjukkan bahwa selisih prediksi terhadap data aktual relatif besar dan model lebih rentan terhadap kesalahan ekstrem. Sementara itu, model Random Forest memproduksi nilai MAE dan RMSE yang secara signifikan lebih kecil, menandakan kemampuan prediksi yang lebih presisi dan stabil. Nilai $R^2$

yang hampir sempurna pada Random Forest (0,9991) juga mengindikasikan bahwa model ini mampu menangkap sebagian besar variasi yang terdapat dalam data, lebih unggul dibandingkan model linier yang memiliki $R^2$ sedikit lebih rendah. Berdasarkan ketiga indikator ini, secara teknis, model Random Forest menunjukkan kemampuan prediksi produksi padi yang lebih unggul dibandingkan metode lainnya.

TABEL V
HASIL PERBANDINGAN AKURASI MODEL

| Model | MAE | RSME | R2 |
|---|---|---|---|
| Linear Regression | 129878.89 | 184311.20 | 0.9947 |
| Random Forest Regression | 40599.94 | 77153.07 | 0.9991 |

Selain evaluasi pada data uji, penelitian ini juga menerapkan Cross Validation (CV) dengan 5 fold untuk menilai konsistensi dan kemampuan generalisasi model. Teknik ini membagi dataset menjadi lima bagian; empat bagian digunakan untuk pelatihan dan satu bagian untuk pengujian secara bergantian hingga seluruh fold terpakai. Pendekatan ini memastikan kemampuan model tidak dipengaruhi hanya oleh satu konfigurasi data, tetapi diuji pada seluruh variasi data.

TABEL VI
HASIL CROSS VALIDATION

| | |
|---|---|
| Cross Validation - Linear Regression | |
| $R^2$ tiap fold | [0.9957, 0.9916, 0.9979, 0.9952, 0.9952] |
| $R^2$ rata-rata | 0.9951 |
| Cross Validation - Random Forest Regression | |
| $R^2$ tiap fold | [0.9956, 0.996, 0.9995, 0.9983, 0.9981] |
| $R^2$ rata-rata | 0.9975 |

## IV. KESIMPULAN

Penelitian ini melakukan pemodelan estimasi produksi padi nasional dengan memanfaatkan variabel luas panen, produktivitas, tahun, dan provinsi berdasarkan data BPS tahun 2018–2024. Hasil eksplorasi data menunjukkan bahwa luas panen merupakan prediktor paling dominan, sementara produktivitas, tahun, dan provinsi berperan sebagai faktor pendukung yang menciptakan variasi spasial dan temporal pada pola produksi. Setelah melewati tahapan preprocessing yang mencakup pembersihan data, transformasi numerik, dan encoding kategorikal, dataset diperoleh dalam bentuk yang konsisten dan siap untuk pemodelan regresi. Pada tahap pemodelan, Regresi Linier Berganda menghasilkan nilai MAE sebesar 159.782,11, RMSE sebesar 238.516,40, dan $R^2$ sebesar 0,9947, yang menunjukkan bahwa model mampu menangkap hubungan dasar antara fitur prediktor dan produksi padi tetapi masih kurang presisi pada area dengan variasi data yang tinggi. Model Random Forest Regression memberikan hasil yang jauh lebih baik, dengan MAE sebesar 40.599,94, RMSE sebesar 77.153,07, dan $R^2$ sebesar 0,9991, menandakan bahwa model ini mampu menangkap pola non-linear dan interaksi antarvariabel secara lebih komprehensif. Hasil visualisasi prediksi vs aktual juga memperlihatkan bahwa Random Forest mampu mengikuti tren aktual secara lebih konsisten dibandingkan pendekatan linier. Pengujian 5-fold Cross Validation semakin menguatkan temuan tersebut. Nilai $R^2$ pada setiap fold untuk Random Forest lebih stabil dan memiliki rata-rata yang lebih besar dibandingkan model linear tersebut, mengindikasikan kemampuan generalisasi yang lebih baik terhadap variasi distribusi data. Sementara itu, Regresi Linier Berganda memperlihatkan fluktuasi performa antar-fold, mengindikasikan sensitivitas terhadap pembagian data dan ketidaksesuaiannya untuk menangani struktur data yang lebih kompleks. Berdasarkan keseluruhan hasil percobaan, hasil penelitian menunjukkan bahwa Random Forest Regression merupakan model terbaik dalam memprediksi produksi padi nasional, ditinjau dari error yang paling rendah, akurasi paling tinggi, konsistensi hasil cross validation, serta kemampuan adaptasi terhadap pola non-linear. Berdasarkan kinerjanya, Random Forest layak dijadikan metode yang lebih terpercaya dibandingkan regresi linier berganda dalam melakukan prediksi produksi padi pada skala nasional.

Berdasarkan hasil eksperimen ang menandakan bahwa algoritma Random Forest Regression memiliki memiliki performa paling optimal pada memodelkan produksi padi nasional, penelitian selanjutnya disarankan untuk memperluas cakupan variabel dengan menambahkan faktor iklim seperti curah hujan dan suhu, penggunaan sarana produksi seperti pupuk dan benih, serta indikator sosial-ekonomi yang berpotensi meningkatkan akurasi model. Selain itu, meskipun Random Forest terbukti unggul dibandingkan regresi linier berganda, pengujian lebih lanjut menggunakan model berbasis ensemble atau deep learning, seperti Gradient Boosting, XGBoost, maupun arsitektur neural network, perlu dilakukan untuk mengevaluasi potensi peningkatan akurasi. Proses validasi juga disarankan untuk diperluas, tidak hanya menggunakan 5-fold cross validation tetapi juga teknik validasi lainnya agar estimasi performa model semakin robust. Visualisasi prediksi dan error yang lebih komprehensif juga dapat digunakan untuk memahami pola ketidaksesuaian model secara lebih mendalam. Dengan pengembangan ini, model prediksi diharapkan dapat digunakan secara lebih efektif sebagai dasar analisis produksi padi dalam mendukung perumusan strategi ketahanan pangan nasional.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Badan Pusat Statistik (BPS) atas penyediaan data resmi yang menjadi dasar terselenggaranya penelitian ini dengan lancar. Penulis juga menghargai dukungan dari Universitas Dian Nuswantoro yang telah memberikan fasilitas, bimbingan akademik, serta lingkungan penelitian yang kondusif. Tanpa adanya kontribusi dari kedua pihak, penelitian ini tidak akan terselesaikan secara optimal.

## DAFTAR PUSTAKA


[1] J. P. Matematika, D. Matematika, T. N. Padilah, and R. I. Adam, "Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang".

[2] R. M. Ikhsanuddin and D. Rusvinasari, "Analisis Pengaruh Luas Area Pertanian Terhadap Prediksi Hasil Pertanian di Kebumen Menggunakan Metode Regresi Linier," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 2, pp. 410–418, Apr. 2025, doi: 10.30591/jpit.v10i2.8471.

[3] A. Rahman, M. Jafar Alamsyah, A. Amiruddin, K. Harun Rasyid, and S. Suhada, *Penerapan Metode Regresi Linear Berganda Untuk Memprediksi Hasil Panen Rumput Laut*, vol. 4, no. 1. 2024.

[4] M. Y. T. Sulistyono, E. S. Pane, E. M. Yuniarno, and M. H. Purnomo, "Correlation Analysis Approach Between Features and Motor Movement Stimulus for Stroke Severity Classification of EEG Signal Based on Time Domain, Frequency Domain, and Signal Decomposition Domain," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 3, Dec. 2024, doi: 10.23887/janapati.v13i3.85550.

[5] J. M. Loban, "Analisis Regresi Faktor-Faktor Yang Mempengaruhi Hasil Produksi Padi Di Indonesia Bagian Barat," Jul. 2023.

[6] D. Noviwiyanah and M. H. Yudhistira, "Pangan Indonesia The Effect Of Paddy Field Area On Indonesian Food Production And Consumption," 2024.

[7] A. Bahtiar, "Prediksi Hasil Panen Padi Tahun 2023 Menggunakan Metode Regresi Linier Di Kabupaten Indramayu," *Jurnal Informatika Terpadu*, vol. 9, no. 1, pp. 18–23, 2023, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/JIT

[8] A. N. A. M. , M. F. F. , F. S. A. L. M. Deris Desmawan, "Dampak Pengalihan Fungsi Lahan Pertanian Menjadi Lahan Permukiman dan Industri Di Kawasan Kabupaten Bekasi," *Bursa: Jurnal Ekonomi dan Bisnis*, vol. 3, no. 2, pp. 116–121, Dec. 2024.

[9] J. Hutahaean and D. Yusup, "Perbandingan Metode Linear Regression, Random Forest & K-Nearest Neighbor Untuk Prediksi Produksi Hasil Panen Padi Di Provinsi Jawa Barat," 2024.

[10] E. Triyanto, H. Sismoro, and A. D. Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi Di Kabupaten Bantul," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 4, no. 2, pp. 66–75, Jul. 2019, doi: 10.36341/rabit.v4i2.666.

[11] D. Nuraini, D. Violina, D. R. Anamisa, B. K. Khotimah, A. Jauhari, and F. A. Mufarroha, "Prediksi Hasil Panen Padi dengan Metode Multiple Linear Regression dan Particle Swarm Optimization untuk Meningkatkan Produksi Padi di Madura," *JUSIFOR : Jurnal Sistem Informasi dan Informatika*, vol. 4, no. 1, pp. 1–8, Jun. 2025, doi: 10.70609/jusifor.v4i1.5857.

[12] M. K. B. Seran, F. Tedy, A. N. Samane, P. Batarius, P. A. Nani, and A. A. J. Sinlae, "Analisis Data Pertanian Tanaman Pangan untuk Memprediksi Hasil Panen di Kabupaten Malaka Menggunakan Metode Multiple Linear Regression," 2024.

[13] 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, Feb. 2024, doi: 10.33480/inti.v18i2.5269.

[14] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *Journal Of Applied Computer Science And Technology (JACOST)*, vol. 4, no. 1, pp. 2723–1453, 2023, doi: 10.52158/jacost.491.

[15] A. Novebrian Maharadja, I. Maulana, and B. Arif Dermawan, "Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi," 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[16] P. Sari Ramadhan and N. Safitri STMIK Triguna Dharma, "Penerapan Data Mining Untuk Mengestimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Deli Serdang," vol. 18, no. SAINTIKOM, pp. 55–61, 2019, [Online]. Available: https://sirusa.bps.go.id/index.php

[17] K. Mahmud Sujon, R. Binti Hassan, Z. Tusnia Towshi, M. A. Othman, M. Abdus Samad, and K. Choi, "When to Use Standardization and Normalization: Empirical Evidence from Machine Learning Models and XAI," *IEEE Access*, vol. 12, pp. 135300–135314, 2024, doi: 10.1109/ACCESS.2024.3462434.

[18] T. O. Hodson, "Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not," Mar. 11, 2022. doi: 10.5194/gmd-2022-64.

[19] M. Fukushige, "Variable Selection and Variable Integration for Categorical Dummy Variables in Regression Analysis," *Annals of Data Science*, 2025, doi: 10.1007/s40745-025-00607-x.

[20] L. Breiman, "Random Forests," 2001.