

JURNAL_UMSIDA_SHERLY_122-1.pdf

by Kim park

Daily Rice Sales Prediction Based on Brand, Price, and Package Size Using Linear Regression and Random Forest Methods

Prediksi Penjualan Beras Harian Berdasarkan Merek, harga, dan Ukuran Kemasan Menggunakan Metode Linear Regression dan Random Forest

Sherly Maziya Rizqi¹⁾, Yulian findawati²⁾, Hamzah Setiawan³⁾, ⁴⁾

^{1,2,3,4)} Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

*Email: yulianfindawati@umsida.com

Abstract. Rice sales volumes fluctuate, necessitating an approach to forecast sales based on historical data. This study aims to apply and compare Linear Regression and Random Forest Regression methods for predicting rice sales volumes and to analyze the impact of attribute selection on the performance of both methods. The study utilizes historical rice sales data comprising the following attributes: Brand, Size, Price per kg, Price per piece, Day, Weekday/Weekend, and Public Holiday. The research stages include data preprocessing, categorical data transformation, splitting data into training and testing sets, correlation analysis and attribute selection, modeling, prediction, and evaluation using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Testing was conducted across three scenarios—utilizing seven, five, and three attributes—based on the ranking of absolute correlation values relative to the target variable. The results indicate that Random Forest Regression yielded lower error rates than Linear Regression across all tested scenarios. The optimal configuration was achieved with Random Forest Regression using five attributes—Brand, Weekday/Weekend, Price per kg, Public Holiday, and Price per piece—resulting in an MAE of 2.3805, an MSE of 17.6392, and an RMSE of 4.1999. Compared to using seven attributes, the use of five attributes with Random Forest Regression reduced MAE by 3.9440%, MSE by 11.0864%, and RMSE by 5.7060%. These findings demonstrate that, for this dataset, attribute selection influences model performance and that utilizing the full set of attributes does not necessarily result in lower prediction error rates.

Keywords - Sales Prediction; Linear Regression; Random Forest Regression; Machine Learning; Rice Sales.

Abstrak. Penjualan beras memiliki jumlah permintaan yang dapat berubah sehingga diperlukan suatu pendekatan untuk memperkirakan jumlah penjualan berdasarkan data historis. Penelitian ini bertujuan untuk menerapkan dan membandingkan metode Linear Regression dan Random Forest Regression dalam memprediksi jumlah penjualan beras serta menganalisis pengaruh pemilihan atribut terhadap kinerja kedua metode. Data yang digunakan merupakan data historis penjualan beras dengan atribut Merek, Ukuran, Harga Per Kg, Harga Per PCS, Hari, Weekday/Weekend, dan Hari Besar. Tahapan penelitian meliputi preprocessing data, transformasi data kategorikal, pembagian data training dan testing, analisis korelasi dan seleksi atribut, pemodelan, prediksi, serta evaluasi menggunakan Mean Absolute Error (MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE). Pengujian dilakukan menggunakan tiga skenario, yaitu tujuh, lima, dan tiga atribut berdasarkan peringkat nilai korelasi absolut terhadap variabel target. Hasil pengujian menunjukkan bahwa Random Forest Regression menghasilkan tingkat kesalahan yang lebih rendah dibandingkan Linear Regression pada seluruh skenario yang diuji. Konfigurasi terbaik diperoleh pada Random Forest Regression menggunakan lima atribut, yaitu Merek, Weekday/Weekend, Harga Per Kg, Hari Besar, dan Harga Per PCS, dengan nilai MAE sebesar 2,3805, MSE sebesar 17,6392, dan RMSE sebesar 4,1999. Dibandingkan penggunaan tujuh atribut, penggunaan lima atribut pada Random Forest Regression menurunkan MAE sebesar 3,9440%, MSE sebesar 11,0864%, dan RMSE sebesar 5,7060%. Hasil tersebut menunjukkan bahwa pada data penelitian ini, pemilihan atribut dapat memengaruhi kinerja model dan penggunaan seluruh atribut tidak selalu menghasilkan tingkat kesalahan prediksi yang lebih rendah.

Kata Kunci - Prediksi Penjualan; Linear Regression; Random Forest Regression; Machine Learning; Penjualan Beras.

I. PENDAHULUAN

Beras merupakan salah satu bahan pangan pokok yang memiliki peran penting dalam memenuhi kebutuhan pangan masyarakat Indonesia. Tingginya tingkat konsumsi beras menunjukkan bahwa komoditas ini masih menjadi salah satu sumber pangan utama yang dibutuhkan masyarakat. Ketergantungan masyarakat terhadap konsumsi beras juga menyebabkan tingkat permintaan beras tetap tinggi [1]. Kondisi tersebut menjadikan ketersediaan beras penting untuk diperhatikan, termasuk dalam kegiatan penjualan untuk memenuhi kebutuhan konsumen.

Dalam kegiatan penjualan, pengelolaan persediaan menjadi salah satu aspek penting untuk menjaga ketersediaan produk sesuai dengan kebutuhan konsumen. Persediaan yang berlebihan dapat menyebabkan

Copyright © Universitas Muhammadiyah Sidoarjo. This preprint is protected by copyright held by Universitas Muhammadiyah Sidoarjo and is distributed under the Creative Commons Attribution License (CC BY). Users may share, distribute, or reproduce the work as long as the original author(s) and copyright holder are credited, and the preprint server is cited per academic standards.

Authors retain the right to publish their work in academic journals where copyright remains with them. Any use, distribution, or reproduction that does not comply with these terms is not permitted.

penumpukan barang dan meningkatkan biaya operasional, sedangkan kekurangan persediaan dapat mengakibatkan hilangnya peluang penjualan. Oleh karena itu, diperlukan pengelolaan persediaan yang tepat untuk membantu menentukan jumlah produk yang perlu disediakan sesuai dengan kebutuhan [2]. Dalam penjualan beras, kondisi tersebut menunjukkan pentingnya memperkirakan jumlah penjualan sebagai salah satu dasar dalam mendukung pengelolaan persediaan.

Data historis penjualan dapat dimanfaatkan sebagai salah satu sumber informasi dalam melakukan prediksi terhadap permintaan di masa mendatang. Melalui pengolahan data historis, pola penjualan yang telah terjadi sebelumnya dapat dianalisis untuk menghasilkan perkiraan yang lebih terarah dan berbasis data [3]. Pemanfaatan data tersebut dapat membantu dalam memahami pola penjualan serta menjadi dasar dalam memperkirakan jumlah produk yang akan terjual pada periode berikutnya.

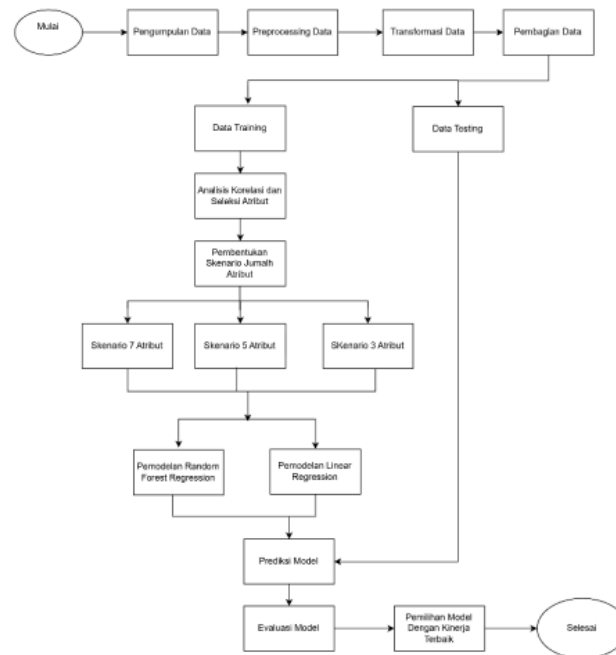
Machine learning merupakan salah satu pendekatan yang dapat digunakan dalam melakukan prediksi dengan mempelajari pola berdasarkan data yang tersedia. Penerapan machine learning pada data penjualan memungkinkan data historis diolah untuk menghasilkan estimasi terhadap penjualan pada periode berikutnya. Penelitian sebelumnya menunjukkan bahwa machine learning dapat diterapkan dalam memprediksi data penjualan maupun transaksi dengan menggunakan algoritma regresi [4]. Selain itu, penggunaan machine learning pada prediksi penjualan retail memungkinkan pola historis dan hubungan antarvariabel dalam data dipelajari sebagai dasar dalam menghasilkan estimasi penjualan [5]. Oleh karena itu, pendekatan machine learning dapat digunakan untuk mendukung proses prediksi penjualan berdasarkan karakteristik data yang digunakan.

Dalam penelitian ini, algoritma yang digunakan untuk melakukan prediksi adalah Linear Regression dan Random Forest Regression. Linear Regression merupakan metode yang memodelkan hubungan antara variabel dependen dan variabel independen untuk menghasilkan nilai prediksi. Metode ini telah diterapkan pada data penjualan dan menunjukkan kemampuan dalam menghasilkan prediksi dengan tingkat kesalahan yang relatif rendah. Sementara itu, Random Forest merupakan metode machine learning berbasis ensemble yang membangun sejumlah decision tree menggunakan subset data dan fitur secara acak, kemudian menggabungkan hasil dari setiap pohon untuk menghasilkan prediksi akhir. Random Forest dapat digunakan untuk permasalahan regresi dengan variabel target kontinu serta memiliki kemampuan dalam mengurangi risiko overfitting melalui proses randomisasi data dan fitur. Oleh karena itu, kedua algoritma tersebut digunakan untuk mengetahui kemampuan masing-masing model dalam memprediksi jumlah penjualan berdasarkan variabel yang digunakan dalam penelitian.

Kinerja model prediksi perlu dievaluasi untuk mengetahui tingkat kesalahan antara nilai hasil prediksi dengan nilai aktual. Pada penelitian ini, evaluasi dilakukan menggunakan Mean Absolute Error (MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE). MAE mengukur rata-rata nilai absolut dari selisih antara nilai aktual dan hasil prediksi, sedangkan MSE menghitung rata-rata kuadrat dari kesalahan prediksi. RMSE merupakan akar dari MSE dan menghasilkan nilai kesalahan dalam satuan yang sama dengan variabel yang diprediksi [8]. Nilai error yang lebih rendah menunjukkan bahwa hasil prediksi model semakin mendekati nilai aktual. Penggunaan ketiga metrik tersebut diharapkan dapat memberikan gambaran mengenai tingkat kesalahan dan performa Linear Regression serta Random Forest Regression dalam memprediksi jumlah penjualan beras.

II. METODE

Penelitian ini menerapkan pendekatan kuantitatif berbasis supervised learning untuk memprediksi jumlah penjualan beras menggunakan metode *Linear Regression* dan *Random Forest Regression*. Adapun alur tahapan penelitian secara keseluruhan ditampilkan pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

A. Pengumpulan Data

Pengumpulan data dalam penelitian ini merupakan data historis penjualan beras harian selama periode 02 Agustus 2025 hingga 31 Mei 2026 dengan jumlah 4.420 data. Dataset mempunyai atribut tanggal, hari, *weekday/weekend*, hari besar, merek, ukuran, harga per kilogram, harga per kemasan, dan jumlah terjual. Dalam pemodelan, variable jumlah terjual digunakan sebagai target prediksi, sedangkan atribut lainnya digunakan sebagai informasi pendukung dalam proses pengolahan data. Contoh dataset penjualan beras ditampilkan pada Tabel 1.

Tabel 1. Contoh Dataset

Tanggal	Hari	Weekday/ Weekend	Hari Besar	Merek	Ukuran	Harga Per Kg	Harga Per Pcs	Jumlah Terjual/PCS
02/08/2025	Sabtu	Weekend	Tidak	Cendrawasih	25 kg	13.400	335.000	21
02/08/2025	Sabtu	Weekend	Tidak	2 Rajawali	25 kg	13.200	330.000	17
02/08/2025	Sabtu	Weekend	Tidak	Labert	25 kg	14.200	355.000	5
02/08/2025	Sabtu	Weekend	Tidak	Rojo Lele	25 kg	13.000	325.000	3
02/08/2025	Sabtu	Weekend	Tidak	Lesung Padi	25 kg	14.800	370.000	1
02/08/2025	Sabtu	Weekend	Tidak	Padamu	25 kg	13.500	337.500	0
02/08/2025	Sabtu	Weekend	Tidak	Polos	25 kg	13.000	325.000	1
02/08/2025	Sabtu	Weekend	Tidak	Cendrawasih	10 Kg	13.600	136.000	9
02/08/2025	Sabtu	Weekend	Tidak	2 Rajawali	10 kg	13.400	134.000	4
02/08/2025	Sabtu	Weekend	Tidak	Labert	10 Kg	14.400	144.000	2

B. Preprocessing Data

Tahap *preprocessing* dilakukan untuk memastikan data yang digunakan dalam proses pemodelan memiliki struktur dan tipe data yang sesuai. Tahap preprocessing dilakukan untuk mempersiapkan data sebelum digunakan dalam proses pemodelan. Proses ini bertujuan untuk meningkatkan kualitas data sehingga data dapat digunakan secara optimal dalam pembentukan model prediksi [6]. Proses ini meliputi pemeriksaan nilai kosong, tipe data, serta konsistensi atribut pada dataset. Dataset awal yang digunakan terdiri dari 4.420 baris data. Hasil pemeriksaan menunjukkan terdapat 119 baris tanpa nilai Jumlah_Terjual karena tidak terdapat aktivitas penjualan pada hari tertentu. Data tersebut tetap menjadi bagian dari dataset awal, tetapi tidak digunakan dalam proses pemodelan karena Jumlah_Terjual merupakan variabel target yang diperlukan dalam proses pelatihan dan pengujian model. Setelah proses preprocessing, diperoleh 4.301 baris data yang digunakan pada tahap selanjutnya.

C. Transformasi Data

Setelah tahap preprocessing, dilakukan transformasi data untuk mengubah atribut kategorikal menjadi bentuk numerik agar dapat digunakan dalam proses pemodelan. Melakukan transformasi data kategorikal menggunakan teknik label encoding, yaitu dengan menetapkan nilai integer pada setiap kategori. Pada penelitian ini, proses encoding diterapkan pada atribut Merek, Hari, *Weekday/Weekend*, dan Hari Besar dengan memberikan kode numerik yang berbeda pada setiap kategori. Kode numerik tersebut digunakan sebagai representasi kategori dalam proses pemodelan.

Sementara itu, atribut Ukuran, Harga Per Kg, dan Harga Per PCS tidak dilakukan encoding karena telah berbentuk numerik. Atribut Ukuran tetap direpresentasikan berdasarkan berat produk dalam kilogram, sedangkan Jumlah Terjual digunakan sebagai variabel target. Hasil transformasi atribut kategorikal ditunjukkan pada Tabel 3.

Tabel 2. Transformasi Atribut Kategorikal

Atribut	Kategori	Kode
Merek	2 rajawali	0
Merek	Cendrawasih	1
Merek	Labert	2
Merek	Lesung padi	3
Merek	Padamu	4
Merek	Polos	5
Merek	Rojo Lele	6
Hari	Kamis	0
Hari	Minggu	1
Hari	Rabu	2
Hari	Sabtu	3
Hari	Selasa	4
Hari	Senin	5
Weekday/Weekend	Weekday	0
Weekend/Weekend	Weekend	1
Hari Besar	Tidak	0
Hari besar	Ya	1

Berdasarkan Tabel 2, setiap kategori direpresentasikan menggunakan kode numerik yang berbeda. Pemberian kode tersebut bertujuan untuk mengubah representasi data kategorikal menjadi numerik sehingga dapat digunakan dalam proses pemodelan. Nilai kode pada masing-masing kategori berfungsi sebagai representasi kategori dalam dataset.

Tabel 3. Contoh Data Setelah Transformasi

Merek	Ukuran	Harga Per Kg	Harga per PCS	Hari	Weekday/Weekend	Hari Besar	Jumlah terjual/pcs
3.0	25	15.200	380.000	4.0	0.0	0.0	4.0
4.0	25	13.800	345.000	4.0	0.0	0.0	1.0
1.0	10	14.100	141.000	4.0	0.0	0.0	9.0
2.0	10	15.500	155.000	4.0	0.0	0.0	3.0
3.0	5	16.200	81.000	2.0	0.0	0.0	2.0

Tabel 3 menunjukkan contoh data setelah proses transformasi, dimana atribut kategorikal telah direpresentasikan dalam bentuk numerik. Data hasil transformasi tersebut selanjutnya digunakan pada tahap pembagian data menjadi data training dan testing.

D. Analisis Korelasi dan Seleksi Atribut

Analisis korelasi digunakan untuk mengetahui tingkat hubungan antara variabel independen dengan variabel dependen. Korelasi Pearson dapat digunakan untuk mengukur kekuatan dan arah hubungan antarvariabel sebelum dilakukan proses pemodelan [7]. Pada penelitian ini, analisis korelasi dilakukan terhadap tujuh atribut dengan variabel target Jumlah Terjual. Nilai korelasi kemudian diubah menjadi nilai absolut dan diurutkan dari nilai terbesar hingga terkecil sebagai dasar pemeringkatan atribut. Berdasarkan hasil pemeringkatan tersebut, dibentuk tiga skenario pengujian, yaitu tujuh atribut, lima atribut dengan nilai korelasi absolut tertinggi, dan tiga atribut dengan nilai korelasi absolut tertinggi. Ketiga skenario selanjutnya diterapkan pada Linear Regression dan Random Forest Regression untuk mengetahui pengaruh jumlah atribut terhadap kinerja prediksi model.

Berdasarkan peringkat nilai korelasi absolut tersebut, dilakukan pengujian menggunakan tiga skenario jumlah atribut. Skenario pertama menggunakan seluruh tujuh atribut, yaitu Merek, Ukuran, Harga Per Kg, Harga Per PCS, Hari, Weekday/Weekend, dan Hari Besar. Skenario kedua menggunakan lima atribut dengan nilai korelasi absolut tertinggi, sedangkan skenario ketiga menggunakan tiga atribut dengan nilai korelasi absolut tertinggi. Pemilihan atribut pada skenario lima dan tiga atribut dilakukan berdasarkan hasil pemeringkatan korelasi yang diperoleh dari data penelitian.

Setiap skenario atribut diterapkan pada Linear Regression dan Random Forest Regression dengan pembagian data training dan testing yang sama. Selanjutnya, hasil prediksi dari masing-masing skenario dievaluasi menggunakan MAE, MSE, dan RMSE. Pengujian ini dilakukan untuk mengetahui pengaruh penggunaan jumlah atribut yang berbeda terhadap tingkat kesalahan prediksi kedua metode serta menentukan konfigurasi yang menghasilkan kinerja terbaik dari skenario yang diuji. Analisis korelasi dilakukan untuk mengetahui hubungan antara masing-masing atribut dengan variabel target Jumlah_Terjual. Nilai korelasi masing-masing atribut kemudian diubah menjadi nilai absolut dan diurutkan dari nilai terbesar hingga terkecil. Penggunaan nilai korelasi absolut bertujuan untuk menentukan peringkat atribut berdasarkan besarnya hubungan terhadap variabel target tanpa membedakan arah hubungan positif maupun negatif.

Berdasarkan peringkat nilai korelasi absolut tersebut, dilakukan pengujian menggunakan tiga skenario jumlah atribut. Skenario pertama menggunakan seluruh tujuh atribut, yaitu Merek, Ukuran, Harga Per Kg, Harga Per PCS, Hari, Weekday/Weekend, dan Hari Besar. Skenario kedua menggunakan lima atribut dengan nilai korelasi absolut tertinggi, sedangkan skenario ketiga menggunakan tiga atribut dengan nilai korelasi absolut tertinggi. Pemilihan atribut pada skenario lima dan tiga atribut dilakukan berdasarkan hasil pemeringkatan korelasi yang diperoleh dari data penelitian.

Setiap skenario atribut diterapkan pada Linear Regression dan Random Forest Regression dengan pembagian data training dan testing yang sama. Selanjutnya, hasil prediksi dari masing-masing skenario dievaluasi menggunakan MAE, MSE, dan RMSE. Pengujian ini dilakukan untuk mengetahui pengaruh penggunaan jumlah atribut yang berbeda terhadap tingkat kesalahan prediksi kedua metode serta menentukan konfigurasi yang menghasilkan kinerja terbaik dari skenario yang diuji.

E. Pembagian Data

Setelah proses transformasi, data dibagi menjadi data *training* dan data *testing* menggunakan metode *random sampling* dengan proporsi 80:20. Dari total 4.300 data yang digunakan dalam proses pemodelan, sebanyak 3.440 data (80%) digunakan sebagai data *training*, sedangkan 860 data (20%) digunakan sebagai data *testing*. Pembagian data dilakukan dengan memisahkan dataset menjadi data training dan data testing. Data training digunakan dalam proses pembentukan model, sedangkan data testing digunakan untuk menguji kemampuan model dalam menghasilkan prediksi terhadap data yang tidak digunakan pada proses pelatihan [8]. Proses pembagian dilakukan secara acak dengan menetapkan *random_state=42* agar pembagian data dapat direproduksi ketika proses dijalankan kembali. Data *training* digunakan untuk melatih model dalam mempelajari pola data penjualan, sedangkan data *testing* digunakan untuk mengevaluasi kemampuan model pada data yang tidak digunakan selama proses pelatihan. Rincian pembagian data ditunjukkan pada Tabel 4.

Tabel 4. Pembagian Data

Jenis Data	Proporsi	Jumlah Data	Keterangan
<i>Training</i>	80%	3.440	Digunakan untuk melatih model.
<i>Testing</i>	20%	861	Digunakan untuk menguji model.
Total	100%	4.301	Data yang digunakan dalam pemodelan.

Pembagian tersebut digunakan secara konsisten pada model *Linear Regression* dan *Random Forest Regression* sehingga kedua model dievaluasi menggunakan data testing yang sama.

F. Linear Regression

Linear Regression merupakan metode yang digunakan untuk membentuk hubungan antara variabel independen dengan variabel dependen. Linear Regression merupakan metode yang dapat digunakan untuk melakukan prediksi berdasarkan hubungan antara variabel yang digunakan dalam penelitian. Metode ini dapat diterapkan pada data historis penjualan untuk menghasilkan estimasi penjualan berdasarkan pola yang terdapat pada data [9]. Pada penelitian ini, model digunakan untuk mempelajari hubungan antara atribut penjualan dengan jumlah beras yang terjual. Hubungan tersebut secara umum dapat dinyatakan melalui persamaan berikut:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (1)$$

Keterangan: Keterangan: y merupakan nilai prediksi, β merupakan parameter regresi, dan x merupakan variabel input.

G. Random Forest Regression

Random Forest Regression merupakan metode ensemble learning yang menggabungkan sejumlah decision tree dalam menghasilkan prediksi. Setiap pohon bekerja sebagai model prediksi secara independen, kemudian hasil prediksi dari seluruh pohon digabungkan untuk memperoleh prediksi akhir [10]. Pada penelitian ini, Random Forest Regression diterapkan untuk mempelajari pola pada data training dan menghasilkan prediksi jumlah penjualan beras pada data testing. Secara umum, hasil prediksi akhir pada regresi dapat dinyatakan sebagai rata-rata prediksi dari seluruh pohon sebagai berikut:

$$y = T1t = 1 \sum Tht(x) \quad (2)$$

Keterangan: y merupakan hasil prediksi, T merupakan jumlah pohon, dan ht(x) merupakan prediksi dari masing-masing pohon.

H. Evaluasi Model

Evaluasi model dilakukan untuk mengetahui tingkat kesalahan prediksi dengan membandingkan nilai aktual dan nilai hasil prediksi. Pada penelitian ini, kinerja model *Linear regression* dan *Random Forest Regression* dievaluasi menggunakan *Mean Absolute error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared error* (RMSE)[11]. Ketiga metrik tersebut digunakan untuk memberikan Gambaran mengenai Tingkat kesalahan dari hasil prediksi model.

1. Mean Absolute Error (MAE)

MAE digunakan untuk mengukur rata-rata kesalahan absolut antara nilai aktual dan nilai hasil prediksi

$$MAE = n1i = 1 \sum n | y_i - y^i | \quad (3)$$

Keterangan: n merupakan jumlah data, y merupakan nilai aktual, dan y^i merupakan nilai prediksi.

2. Mean Squared Error (MSE)

MSE digunakan untuk mengukur rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi sehingga kesalahan yang besar memperoleh bobot yang lebih besar.

$$MSE = n1i = 1 \sum n (y_i - y^i)^2 \quad (4)$$

Keterangan: Keterangan: n merupakan jumlah data, y merupakan nilai aktual, dan y^i merupakan nilai prediksi.

3. Root Mean Squared Error (RMSE)

RMSE merupakan akar kuadrat dari MSE yang digunakan untuk mengukur tingkat kesalahan prediksi dalam satuan yang sama dengan data actual.

$$RMSE = n1i = 1 \sum n (y_i - y^i)^2 \quad (5)$$

Keterangan: n merupakan jumlah data, y merupakan nilai aktual, dan y^i merupakan nilai prediksi.

III. HASIL DAN PEMBAHASAN

A. Visualisasi Data

Data yang digunakan dalam penelitian ini merupakan data historis penjualan beras harian. Dataset awal terdiri atas 4.420 baris data. Setelah dilakukan preprocessing pada variabel target Jumlah Terjual, terdapat 119 baris data yang tidak memiliki nilai target, sehingga diperoleh 4.301 baris data yang digunakan dalam proses pemodelan.

Untuk memberikan gambaran umum mengenai karakteristik data numerik yang digunakan dalam penelitian, dilakukan analisis statistik deskriptif terhadap atribut Ukuran, Harga Per Kg, Harga Per PCS, dan Jumlah Terjual. Statistik deskriptif meliputi nilai minimum, maksimum, dan rata-rata sebagaimana ditunjukkan pada Tabel 5.

Tabel 5. Statistik Deskriptif Data Penjualan Beras

Atribut	Minimum	Maksimum	Rata-rata
Ukuran	3 Kg	25 Kg	14 Kg
Harga Per Kg	13.000	16.400	14.511,18 Kg
Harga Per PCS	43.000	400.000	199.213,900
Jumlah Terjual	0 PCS	76 PCS	4,5 PCS

Berdasarkan Tabel 5, ukuran kemasan beras berada pada rentang 3 kg hingga 25 kg dengan rata-rata 14 kg. Harga per kilogram memiliki nilai minimum Rp13.000 dan maksimum Rp16.400, dengan rata-rata sebesar Rp14.511,18. Harga per kemasan berada pada rentang Rp43.000 hingga Rp400.000, dengan rata-rata sebesar Rp199.213,90. Sementara itu, jumlah penjualan memiliki nilai minimum 0 unit, maksimum 76 unit, dan rata-rata sebesar 4,50 unit

B. Hasil Prediksi Linear Regression

Linear Regression digunakan untuk menghasilkan prediksi jumlah penjualan beras pada data pengujian. Pada bagian ini, hasil prediksi ditampilkan menggunakan skenario lima atribut, yaitu Merek, Weekday/Weekend, Harga Per Kg, Hari Besar, dan Harga Per PCS. Beberapa hasil prediksi Linear Regression pada data pengujian ditunjukkan pada Tabel 6.

Tabel 6. Hasil Prediksi Linear Regression Menggunakan Lima Atribut

Tanggal	Merek	Ukuran (Kg)	Aktual	Prediksi
23/08/2025	Rojo Lele	25	5	3.4627
24/05/2026	Cendrawasih	5	4	5.7207
05/11/2025	Lesung Padi	25	1	3.2017
15/02/2026	Polos	25	0	3.7628
09/11/2025	Rojo Lele	25	8	3.2437

Berdasarkan Tabel 6, hasil prediksi Linear Regression menunjukkan adanya perbedaan antara nilai aktual dan nilai prediksi pada data pengujian. Pada penjualan Rojo Lele tanggal 23 Agustus 2025 dengan nilai aktual 5 unit, model menghasilkan prediksi sebesar 3,4627 unit. Pada Cendrawasih tanggal 24 Mei 2026 dengan nilai aktual 4 unit, hasil prediksi sebesar 5,7207 unit. Sementara itu, perbedaan yang lebih besar terlihat pada Rojo Lele tanggal 9 November 2025, yaitu nilai aktual sebesar 8 unit sedangkan model menghasilkan prediksi sebesar 3,2437 unit. Hasil tersebut menunjukkan bahwa kemampuan model dalam mendekati nilai aktual berbeda pada setiap data pengujian. Oleh karena itu, kinerja Linear Regression secara keseluruhan selanjutnya dianalisis menggunakan MAE, MSE, dan RMSE pada bagian evaluasi dan perbandingan model.

C. Hasil Prediksi Random Forest Regression

Random Forest Regression digunakan untuk menghasilkan prediksi jumlah penjualan beras pada data pengujian dengan menggunakan skenario lima atribut, yaitu Merek, Weekday/Weekend, Harga Per Kg, Hari Besar, dan Harga Per PCS. Beberapa hasil prediksi Random Forest Regression pada data pengujian ditunjukkan pada Tabel 7.

Tabel 7. Hasil prediksi Random Forest Regression Menggunakan Lima Atribut

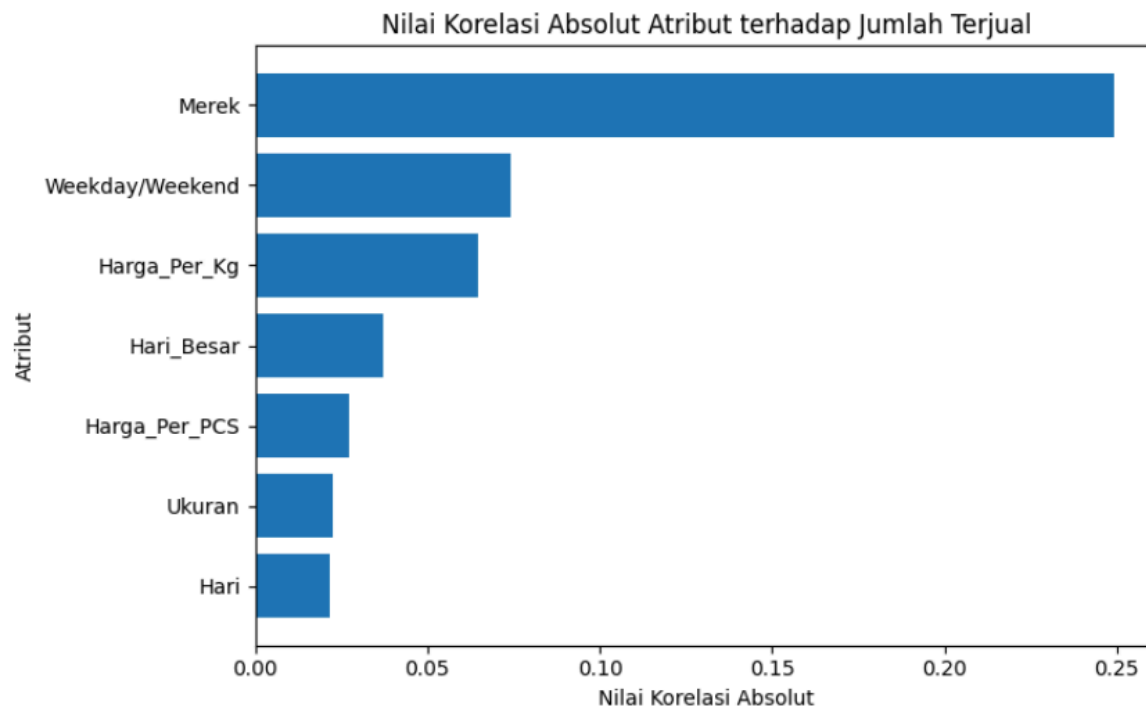
Tanggal	Merek	Ukuran(Kg)	Aktual	Prediksi
23/08/2025	Rojo Lele	25	5	2.3193
24/05/2026	Cendrawasih	5	4	4.0834
05/11/2025	Lesung Padi	25	1	2.2136
15/02/2026	Polos	25	0	1.6886
09/11/2025	Rojo Lele	25	8	3.4131

Berdasarkan Tabel 7, Random Forest Regression menghasilkan nilai prediksi yang bervariasi terhadap nilai aktual pada masing-masing data pengujian. Pada data Cendrawasih tanggal 24 Mei 2026, nilai aktual sebesar 4 unit memperoleh hasil prediksi sebesar 4,0834 unit sehingga menunjukkan selisih yang relatif kecil. Pada data Polos tanggal 15 Februari 2026 dengan nilai aktual 0 unit, model menghasilkan prediksi sebesar 1,6886 unit. Sementara itu, pada data Rojo Lele tanggal 9 November 2025 dengan nilai aktual 8 unit, hasil prediksi sebesar 3,4131 unit menunjukkan adanya selisih yang lebih besar. Perbedaan tersebut menunjukkan bahwa hasil prediksi model tidak

selalu memiliki tingkat kedekatan yang sama terhadap nilai aktual pada setiap data pengujian. Kinerja Random Forest Regression secara keseluruhan selanjutnya dianalisis menggunakan MAE, MSE, dan RMSE pada bagian evaluasi dan perbandingan model.

D. Hasil analisis Korelasi dan Seleksi Atribut

Setelah data melalui tahap preprocessing dan transformasi, dilakukan analisis korelasi antara atribut yang digunakan dalam pemodelan dengan variabel target Jumlah_Terjual. Analisis ini digunakan untuk memperoleh peringkat atribut berdasarkan nilai korelasi absolut terhadap variabel target. Hasil analisis korelasi ditunjukkan pada Gambar 2.



Gambar 2. Nilai Korelasi Absolut Atribut Terhadap Jumlah Terjual

Berdasarkan Gambar 2, atribut Merek memperoleh nilai korelasi absolut tertinggi sebesar 0,2492, diikuti oleh Weekday/Weekend sebesar 0,0740 dan Harga Per Kg sebesar 0,0647. Selanjutnya, atribut Hari Besar memperoleh nilai sebesar 0,0371, Harga Per PCS sebesar 0,0274, Ukuran sebesar 0,0226, dan Hari sebesar 0,0219. Hasil tersebut menunjukkan urutan nilai korelasi absolut masing-masing atribut terhadap variabel target pada data penelitian.

Berdasarkan urutan tersebut, dibentuk tiga skenario atribut yang selanjutnya digunakan dalam pengujian kedua metode. Skenario tujuh atribut menggunakan seluruh atribut penelitian. Skenario lima atribut menggunakan Merek, Weekday/Weekend, Harga Per Kg, Hari Besar, dan Harga Per PCS, sedangkan skenario tiga atribut menggunakan Merek, Weekday/Weekend, dan Harga Per Kg. Ketiga skenario tersebut kemudian diterapkan pada Linear Regression dan Random Forest Regression untuk membandingkan kinerja model berdasarkan nilai MAE, MSE, dan RMSE.

E. Evaluasi dan Perbandingan Model pada Skenario Atribut

Evaluasi dilakukan terhadap hasil prediksi Linear Regression dan Random Forest Regression pada tiga skenario jumlah atribut, yaitu 7, 5, dan 3 atribut. Evaluasi menggunakan MAE, MSE, dan RMSE untuk membandingkan tingkat kesalahan prediksi pada masing-masing skenario. Hasil evaluasi keseluruhan ditunjukkan pada Tabel 8.

Tabel 8. Hasil Evaluasi Model Berdasarkan Skenario Jumlah Atribut

Jumlah Atribut	Metode	MAE	MSE	RMSE
7 Atribut	Linear Regression	2,7726	24,5809	4,9506
7 Atribut	Random Forest Regression	2,4782	19,8386	4,4541
5 Atribut	Linear Regression	2,7602	24,5932	4,9591
5 Atribut	Random Forest Regression	2,3805	17,6392	4,1999
3 Atribut	Linear Regression	2,7460	24,6536	4,9652
3 Atribut	Random Forest Regression	2,3840	17,8679	4,2270

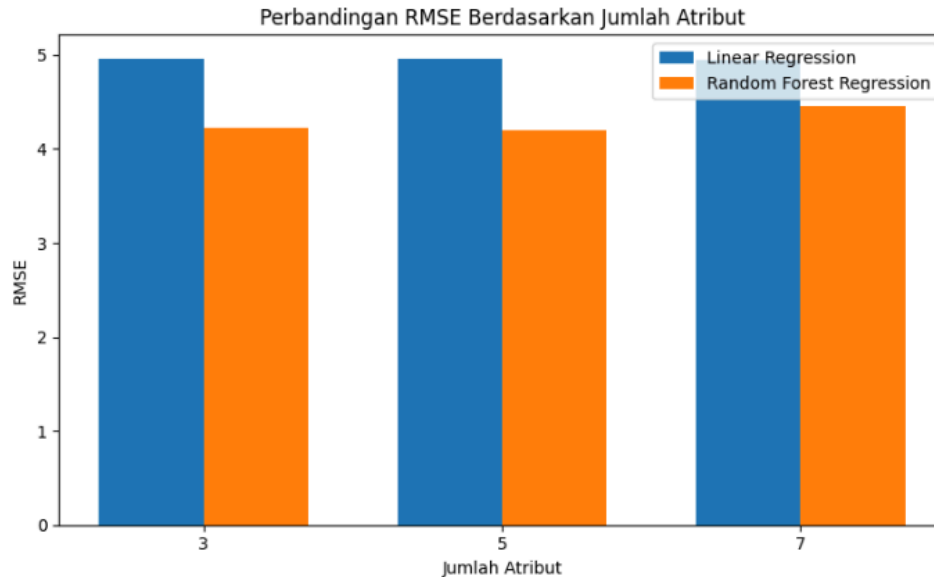
Berdasarkan Tabel 8, Random Forest Regression menghasilkan nilai MAE, MSE, dan RMSE yang lebih rendah dibandingkan Linear Regression pada seluruh skenario jumlah atribut. Pada penggunaan tujuh atribut, Random Forest Regression memperoleh MAE sebesar 2,4782, MSE sebesar 19,8386, dan RMSE sebesar 4,4541, sedangkan Linear Regression memperoleh MAE sebesar 2,7726, MSE sebesar 24,5089, dan RMSE sebesar 4,9506.

Perubahan jumlah atribut memberikan pengaruh yang berbeda terhadap kedua metode. Pada Linear Regression, pengurangan jumlah atribut dari tujuh menjadi lima dan tiga atribut menghasilkan penurunan nilai MAE dari 2,7726 menjadi 2,7602 dan 2,7460. Namun, nilai MSE meningkat dari 24,5089 menjadi 24,5932 dan 24,6536, sedangkan RMSE meningkat dari 4,9506 menjadi 4,9591 dan 4,9652. Hasil tersebut menunjukkan bahwa pengurangan jumlah atribut tidak memberikan peningkatan kinerja yang konsisten pada Linear Regression.

Berbeda dengan Linear Regression, Random Forest Regression menunjukkan penurunan pada seluruh metrik evaluasi ketika jumlah atribut dikurangi dari tujuh menjadi lima. Nilai MAE menurun dari 2,4782 menjadi 2,3805, MSE dari 19,8386 menjadi 17,6392, dan RMSE dari 4,4541 menjadi 4,1999. Pada penggunaan tiga atribut, nilai MAE,

MSE, dan RMSE masing-masing sebesar 2,3840, 17,8679, dan 4,2270. Meskipun hasil tiga atribut mendekati hasil lima atribut, skenario lima atribut tetap menghasilkan nilai kesalahan terendah pada ketiga metrik evaluasi.

Untuk memperjelas perubahan kinerja kedua metode pada masing-masing skenario atribut, perbandingan nilai RMSE ditampilkan pada Gambar 3



Gambar 4. Perbandingan RMSE Linear Regression dan Random Forest Regression Berdasarkan Jumlah Atribut.

Berdasarkan keseluruhan hasil pengujian, Random Forest Regression dengan lima atribut merupakan konfigurasi dengan tingkat kesalahan prediksi paling rendah dari seluruh skenario yang diuji, dengan nilai MAE sebesar 2,3805, MSE sebesar 17,6392, dan RMSE sebesar 4,1999. Hasil tersebut menunjukkan bahwa pada data pengujian penelitian ini, penggunaan seluruh atribut tidak menghasilkan kinerja terbaik pada Random Forest Regression. Penggunaan lima atribut dengan nilai korelasi absolut tertinggi menghasilkan nilai kesalahan yang lebih rendah dibandingkan penggunaan tujuh maupun tiga atribut.

IV. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, metode Linear Regression dan Random Forest Regression dapat diterapkan untuk memprediksi jumlah penjualan beras menggunakan data historis penjualan. Pengujian dilakukan menggunakan tiga skenario jumlah atribut, yaitu tujuh, lima, dan tiga atribut berdasarkan peringkat nilai korelasi absolut terhadap variabel target. Hasil evaluasi menunjukkan bahwa Random Forest Regression menghasilkan nilai MAE, MSE, dan RMSE yang lebih rendah dibandingkan Linear Regression pada seluruh skenario yang diuji. Perubahan jumlah atribut memberikan pengaruh yang berbeda terhadap kinerja kedua metode. Pada Linear Regression, pengurangan jumlah atribut tidak menghasilkan peningkatan kinerja yang konsisten pada seluruh metrik evaluasi. Sementara itu, pada Random Forest Regression, penggunaan lima atribut menghasilkan tingkat kesalahan yang lebih rendah dibandingkan penggunaan tujuh maupun tiga atribut. Konfigurasi terbaik dari seluruh skenario yang diuji diperoleh pada Random Forest Regression dengan lima atribut, yaitu Merek, Weekday/Weekend, Harga Per Kg, Hari Besar, dan Harga Per PCS, dengan nilai MAE sebesar 2,3805, MSE sebesar 17,6392, dan RMSE sebesar 4,1999. Dibandingkan penggunaan tujuh atribut, skenario lima atribut pada Random Forest Regression menghasilkan penurunan MAE sebesar 3,9440%, MSE sebesar 11,0864%, dan RMSE sebesar 5,7060%. Hasil tersebut menunjukkan bahwa pada data penelitian ini, penggunaan seluruh atribut tidak selalu menghasilkan tingkat kesalahan prediksi yang lebih rendah dan pemilihan atribut dapat memengaruhi kinerja model. Penelitian selanjutnya dapat mempertimbangkan penggunaan variabel lain yang relevan terhadap perubahan jumlah penjualan serta melakukan optimasi parameter model untuk memperoleh kinerja prediksi yang lebih baik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Muhammadiyah Sidoarjo, khususnya Program Studi Informatika Fakultas Sains dan Teknologi, yang telah memberikan dukungan dan fasilitas dalam pelaksanaan penelitian ini. Penulis juga menyampaikan terima kasih kepada orang tua yang senantiasa kebersamaian, mendoakan, memberikan dukungan, serta motivasi selama proses penyusunan penelitian ini hingga dapat terselesaikan dengan baik.

REFERENSI

- [1] L. Habibah, A. Futri, A. P. Khuzaeri, F. Shidqi, and W. A. Winata, "Bursa : Jurnal Ekonomi dan Bisnis Beras Sebagai Makanan Pokok : Faktor Penyebab Ketergantungan Dan," *J. Ekon. Bisnis*, vol. 3, no. 2, pp. 110–114, 2024, [Online]. Available: <https://jurnal.risetilmiah.ac.id/index.php/jeb%0ABeras>
- [2] N. B. Hilalia, "As-Syirkah: Islamic Economics & Finacial Journal Optimalkan Pengelolaan Persediaan Untuk Mengurangi Kerugian: Strategi Menghadapi Permasalahan Persediaan Rusak," *As Syirkah Islam. Econ. Finacial J.*, vol. Volume 3 N, no. Optimalkan Pengelolaan Persediaan Untuk Mengurangi Kerugian : Strategi Menghadapi Persediaan Rusak, pp. 767–778, 2024, doi: 10.56672/assyrakah.v3i2.200.
- [3] A. Asrul *et al.*, "Pemanfaatan Big Data Analytics dalam Proses Manajemen Teknologi untuk Prediksi Permintaan Pasar," *J. Minfo Polgan*, vol. 13, no. 2, pp. 2433–2438, 2025, doi: 10.33395/jmp.v13i2.14516.
- [4] Frira Sesilia, Viktor Handrianus Pranatawijaya, and Ressa Priskila, "Machine Learning untuk Memprediksi Jumlah Penjualan, Stok dan Jumlah Tanam Hasil Pertanian Hidroponik," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 4, no. 1, pp. 222–233, 2024, doi: 10.24002/konstelasi.v4i1.9055.
- [5] J. Ilmu, T. Informasi, V. No, and J. Hal, "STUDI PERBANDINGAN ALGORITMA MACHINE LEARNING," vol. 1, no. 1, pp. 22–28, 2026.
- [6] Ilham Amansyah, Jamaludin Indra, Euis Nurlaesari, and Ayu Ratna Juwita, "Prediksi Penjualan Kendaraan Menggunakan Regresi Linear: Studi Kasus pada Industri Otomotif di Indonesia," *J. Soc. Sci. Res.*, vol. Volume 4 Nomor 4, p. Page 1199-1216, 2024.
- [7] S. A. Pradana *et al.*, "Penjualan Produk Berbasis Web Pada Umkm Toko," vol. 4, no. 4, 2026.
- [8] F. Nur Farida, A. Faqih, and S. Eka Permana, "Penerapan Model Prediksi Penjualan Pada Usaha Rumah Makan Menggunakan Algoritma Random Forest," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 4, pp. 5895–5902, 2025, doi: 10.36040/jati.v9i4.13912.
- [9] H. Wijaya, D. Pramana Hostiadi, and E. Triandini, "Meningkatkan Prediksi Penjualan Retail Xyz Dengan Teknik Optimasi Random Search Pada Model Xgboost," *□ Spinter*, vol. 1, no. 2, p. 2024, 2024.
- [10] R. Hidayat *et al.*, "Implementasi Algoritma Random Forest Regression Untuk Memprediksi Penjualan Produksi di Supermarket," *Simkom*, vol. 10, no. 1, pp. 101–109, 2025, doi: 10.51717/simkom.v10i1.703.
- [11] M. Hardika, Martono, Arif Rinaldy Dikananda, "Meningkatkan Model Prediksi Penjualan Pada Toko Devanjayaban," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 13, no. 2, pp. 913–921, 2025.

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © Universitas Muhammadiyah Sidoarjo. This preprint is protected by copyright held by Universitas Muhammadiyah Sidoarjo and is distributed under the Creative Commons Attribution License (CC BY). Users may share, distribute, or reproduce the work as long as the original author(s) and copyright holder are credited, and the preprint server is cited per academic standards.

Authors retain the right to publish their work in academic journals where copyright remains with them. Any use, distribution, or reproduction that does not comply with these terms is not permitted.

ORIGINALITY REPORT

18%
SIMILARITY INDEX

19%
INTERNET SOURCES

22%
PUBLICATIONS

18%
STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to International University of Tourism 'Silk Road' Student Paper	16%
2	Submitted to Universitas Pamulang Student Paper	1%
3	etheses.uin-malang.ac.id Internet Source	1%
4	Rani Tasya Samosir, Suroso Suroso, Irma Salamah. "Prediksi Waktu Tunggu Pada Sistem Antrean Pasien Berbasis Mobile Menggunakan Algoritma Random Forest", Jurnal Minfo Polgan, 2026 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		