

Penerapan Metode Regresi Linier Berganda Untuk Memprediksi Panen Kelapa Sawit

Hermansyah^{1*}, Asrul Abdullah², Putri Yuli Utami³

Teknik Informatika, Universitas Muhammadiyah Pontianak, Pontianak, Indonesia

*e-mail *Corresponding Author*: hermansyah6452@gmail.com

Abstract

Perkebunan Nusantara XIII Kebun Rimba Belian requires effective planning and strategies to increase production yields. This research aims to apply the multiple linear regression method to predict palm oil production results in these plantations. The research uses a multiple linear regression method by observing patterns of increase or decrease in production results and predicting production results in the next few months. The results of the research show that the multiple linear regression method is effectively used to predict oil palm harvest at Perkebunan Nusantara XIII Kebun Rimba Belian. Analysis shows that the model has a high level of accuracy, with a Root Mean Squared Error (RMSE) value of 0.0698 and an R-squared (R²) Score of 0.9306. This indicates that the model has good abilities in predicting target values and explaining data variations well. As a result, this model can be a useful tool in planning plant care and pest control activities to increase oil palm production yields.

Keywords: *Data mining; Palm oil; Production prediction; Multiple linear regression.*

Abstrak

Perkebunan Nusantara XIII Kebun Rimba Belian memerlukan perencanaan dan strategi yang efektif untuk meningkatkan hasil produksi. Penelitian ini bertujuan untuk menerapkan metode regresi linier berganda untuk memprediksi hasil produksi kelapa sawit di perkebunan tersebut. Penelitian menggunakan metode regresi linier berganda dengan mengamati pola peningkatan atau penurunan hasil produksi dan memprediksi hasil produksi beberapa bulan ke depan. Hasil penelitian menunjukkan bahwa metode regresi linier berganda efektif digunakan untuk memprediksi panen kelapa sawit di Perkebunan Nusantara XIII Kebun Rimba Belian. Analisis menunjukkan bahwa model memiliki tingkat akurasi yang tinggi, dengan nilai *Root Mean Squared Error* (RMSE) sebesar 0.0698 dan R-squared (R²) Score sebesar 0.9306. Hal ini menandakan bahwa model memiliki kemampuan baik dalam memprediksi nilai target dan menjelaskan variasi data dengan baik. Sebagai hasilnya, model ini dapat menjadi alat yang berguna dalam merencanakan kegiatan perawatan tanaman dan pengendalian hama untuk meningkatkan hasil produksi kelapa sawit.

Kata Kunci: *Data mining; Kelapa sawit; Prediksi produksi; regresi linier berganda.*

1. Pendahuluan

Industri kelapa sawit merupakan sektor yang sangat penting dalam perekonomian Indonesia karena mendatangkan sebagian besar pendapatan pemerintah dan menciptakan lapangan kerja bagi masyarakat. Perencanaan dan pengelolaan yang tepat diperlukan untuk memaksimalkan produksi minyak sawit. Perkebunan Nusantara XIII atau disingkat PTPN XIII Kebun Rimba Belian yang bergerak di bidang perkebunan kelapa sawit memerlukan perencanaan dan strategi untuk meningkatkan hasil produksi kelapa sawit yang lebih baik. Salah satu cara untuk meningkatkan hasil adalah dengan memperbanyak penanaman atau mengoptimalkan perawatan tanaman sawit yang sudah berproduksi, seperti pemupukan, pembersihan lahan dari tanaman pengganggu, dan pembasmian hama [1].

Melalui analisis situasi, terlihat bahwa terdapat kesenjangan antara kondisi aktual produksi kelapa sawit di PTPN XIII Kebun Rimba Belian dengan kondisi ideal yang diharapkan. Gap ini menciptakan tantangan terukur dalam upaya meningkatkan hasil produksi. Merinci perbedaan ini, penelitian ini mampu mengidentifikasi masalah yang perlu diatasi melalui solusi yang dapat diukur dan terukur. Dalam merumuskan rencana atau konsep solusi, penelitian ini

didasari oleh konsep regresi linier berganda, yang telah terbukti efektif dalam prediksi hasil produksi berdasarkan variabel-variabel tertentu.

Pada penelitian yang dilakukan oleh [2] menggunakan metode regresi linier untuk memprediksi hasil panen kelapa sawit dengan nilai kecocokan dari model regresi linier berganda sebesar 93,52% dan nilai akurasi dari *Root Mean Square Error* (RMSE) sebesar 0,468. Penelitian yang berkaitan dengan penelitian ini yaitu penerapan algoritma *linear regression* untuk prediksi hasil panen tanaman padi oleh [3] dengan hasil nilai rata-rata akurasi RMSE yang dihasilkan, sebesar 0,432, menunjukkan bahwa variasi nilai yang dihasilkan oleh suatu model prakiraan mendekati akurat. Referensi tersebut akan digunakan untuk memberikan landasan teoritis yang kokoh, mengarah pada solusi yang rasional dan terukur. Oleh karena itu, metode ini diharapkan dapat memberikan kontribusi positif terhadap peningkatan hasil produksi kelapa sawit.

Tujuan utama penelitian ini adalah Memperoleh prediksi produksi kelapa sawit menggunakan metode regresi linier berganda berdasarkan data hasil produksi kelapa sawit. Mengetahui nilai terendah dari *Root Mean Squared Error* menggunakan metode Regresi Linier Berganda dalam melakukan prediksi hasil produksi. Adapun manfaat penelitian yang dilakukan yaitu Mendapatkan informasi gambaran dari prediksi hasil produksi kelapa sawit untuk masa yang akan datang. Dapat dijadikan acuan dalam menentukan strategi untuk meningkatkan hasil produksi kelapa sawit. Memahami metode Regresi Linier Berganda untuk memprediksi hasil produksi kelapa sawit di PTPN 13 Kebun Rimba Belian.

2. Tinjauan Pustaka

Penelitian-Penelitian Terdahulu dalam Prediksi Panen Hasil Pertanian/Perkebunan Sejumlah penelitian terdahulu telah dilakukan untuk mengeksplorasi dan mengatasi masalah prediksi panen hasil pertanian atau perkebunan. Penelitian-penelitian ini memberikan wawasan dan kontribusi yang signifikan dalam menghadapi tantangan serupa yang dihadapi oleh riset ini. Salah satu penelitian yang relevan adalah Penelitian [4] dengan judul penerapan data mining dalam prediksi hasil produksi kelapa sawit PT Borneo Ketapang Indah menggunakan metode regresi linier. Hasil ramalan jumlah produksi minyak sawit di PT Borneo Ketapang Indah untuk tiga bulan ke depan pada bulan April adalah: 350862.531, Mei: 304206.794, dan pada bulan Juni menjadi: 389668.603. Dari ketiga variabel yang digunakan untuk memprediksi jumlah produksi kelapa sawit meliputi luas lahan (X_1), curah hujan (X_2), dan pupuk (X_3). Kemudian subvariabel curah hujan dan negara memiliki pengaruh terbesar terhadap jumlah produksi minyak sawit ketika nilai koefisien luas (X_1) untuk masing-masing negara adalah 0,646 dan curah hujan (X_2) adalah 0,530. Dengan bantuan perhitungan metode regresi linier dapat menjadi solusi bagi perusahaan untuk mengatasi masalah ketidakstabilan produksi minyak sawit.

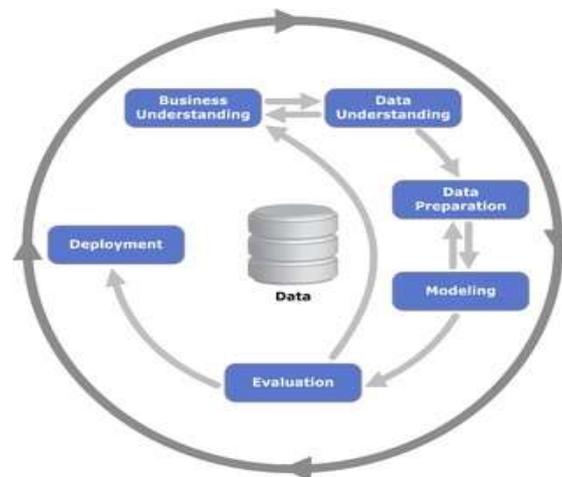
Selanjutnya, Prediksi produksi jagung menggunakan algoritma apriori dan regresi linier berganda (studi kasus : dinas pertanian kabupaten dompu). Hasil perhitungan teknik regresi linier berganda untuk prediksi jagung menggunakan uji MAD adalah 54, MSE adalah 231372 dan RMSE adalah 481, dan hasil persamaan regresi linier berganda adalah $Y = -70,860 + -0,505 x_1 + 7,069 x_2 + -4,349 x_3$. Perhitungan dengan menggunakan teknik asosiasi menghasilkan 2 rule dengan nilai support minimal 10 dengan tingkat kepastian 70% yaitu *IF BUY Gandasil THEN BUY Ricestar with confidence 100%* [5].

Penelitian lain berjudul prediksi jumlah produksi kelapa sawit dengan regresi linier berganda. Analisis uji t parsial untuk buah segar (X_1) berpengaruh nyata terhadap jumlah produksi kelapa sawit (Y), catatan kehadiran panen (X_2) berpengaruh tidak nyata terhadap jumlah produksi kelapa sawit (Y). Secara simultan (bersama-sama) buah anggur segar (X_1) dan pemanen (X_2) berpengaruh nyata terhadap total produksi minyak sawit (Y). Didapatkan pula MAD 43.636.311,67, MSE (*Mean Square Error*) 1.916.114.651.658.140, RMSE 43.773.446,88 dan MAPE 78,59% [6].

Penelitian lain yang relevan adalah tentang penerapan algoritma linear regression untuk prediksi hasil panen tanaman padi. Pada penelitian ini menggunakan beberapa variabel yaitu luas lahan, jumlah bibit, pupuk urea dan pupuk NPK phonska. Level dibuat dari total data 300 instance kecocokan model regresi linier berganda adalah 94,51%. Ini berarti bahwa hingga 94,51% dari variasi nilai yang diperoleh disebabkan oleh variabel independen yang diukur, yaitu. H. luas lahan, varietas bibit, bibit, pupuk urea dan pupuk NPK Phonska. Selebihnya, 5,49% dipengaruhi oleh variabel lain yang tidak diukur dalam jurnal ini. Hasil rata-rata skor akurasi RMSE adalah 0,432. Hal ini menunjukkan bahwa variasi nilai yang dihasilkan oleh model prediksi yang dihasilkan hampir akurat[7].

Perbedaan signifikan dalam konsep dan metode yang digunakan dibandingkan dengan penelitian-penelitian terdahulu, terutama riset ini menonjol dengan menggunakan pendekatan regresi linier berganda untuk memprediksi hasil panen produksi kelapa sawit secara simultan. Penelitian ini memperkaya konsep dengan memasukkan variabel-variabel yang lebih komprehensif, termasuk data bulan, curah hujan, jumlah pohon, luas lahan, dan pupuk. Pendekatan ini diharapkan dapat memberikan prediksi yang lebih akurat dan mengatasi ketidakstabilan produksi yang mungkin terjadi. Selain itu, penelitian saat ini memanfaatkan *Root Mean Squared Error* (RMSE) sebagai indikator kinerja utama. Penggunaan RMSE memberikan gambaran yang lebih jelas tentang sejauh mana model mampu memprediksi hasil panen dengan akurat, memberikan nilai yang optimal untuk evaluasi model. Dengan demikian, state of the art riset ini tidak hanya memperkaya konsep tetapi juga meningkatkan metodologi prediksi panen hasil pertanian dan produksi kelapa sawit dengan pendekatan yang lebih holistik dan evaluatif.

3. Metodologi



Gambar 1 Tahapan CRISP-DM

Perancangan sistem penerapan metode regresi linier berganda untuk memprediksi panen kelapa sawit dengan menggunakan *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) adalah suatu proses yang melibatkan langkah-langkah berikut [8]:

- 1) *Understanding the Business Problem* (Memahami Permasalahan Bisnis)
Identifikasi tujuan dari penelitian ini, yaitu memprediksi hasil panen kelapa sawit. Hasil prediksi akan digunakan untuk mendapatkan informasi gambaran dari prediksi hasil produksi kelapa sawit untuk masa yang akan datang, sehingga dapat dijadikan acuan dalam menentukan strategi untuk meningkatkan hasil produksi kelapa sawit. Tujuan lain bagi peneliti untuk lebih memahami metode Regression Linear berganda untuk memprediksi hasil produksi kelapa sawit di PTPN 13 Kebun Rimba Belian. Data yang diambil data hasil produksi kelapa sawit tahun 2019-2022.
- 2) *Data Understanding* (Memahami Data)
Kumpulkan data yang relevan untuk prediksi hasil panen kelapa sawit, seperti data curah hujan, jumlah pohon, luas lahan, dan pupuk. Lakukan eksplorasi data untuk memahami karakteristik dan kualitas data, seperti melihat *statistika deskriptif*, *visualisasi data*, dan analisis *korelasi*.

```
print('Statistika Deskriptif')
display(df.describe())
```

Dari Code digunakan untuk menghitung beberapa angka penting tentang data, seperti angka rata-rata, angka terkecil, angka terbesar, dan lainnya. Ini membantu untuk memahami bagaimana data itu tersebar dan apakah ada nilai yang aneh. Ini adalah cara cepat untuk memahami data sebelum mulai menganalisisnya dengan lebih dalam.

```
# Data Visualization
1. fig, axes = plt.subplots(2, 4, figsize=(15, 10))
2. axes = axes.flatten()
3. features = ['Afdeling', 'Tahun', 'Bulan', 'Luas', 'Jumlah Pohon', 'curah hujan', 'NPK', 'Dolomite']
4. for i, feature in enumerate(features):
5. axes[i].scatter(df[feature], df['Produksi'], color='blue', alpha=0.5)
6. axes[i].set_xlabel(feature)
7. axes[i].set_ylabel('Produksi')
8. axes[i].set_title(f'{feature} vs Produksi')
9. plt.tight_layout()
10. plt.show()
```

Tujuan dari Code adalah untuk melihat apakah ada pola atau hubungan antara setiap fitur dengan 'Produksi'. Jika melihat pola tertentu, dapat mengambil keputusan lebih lanjut.

```
correlations = df.corr()
print(correlations['Produksi'])
```

Melalui analisis korelasi seperti pada Code, kita dapat melihat apakah ada korelasi antara fitur Afdeling, Tahun, Bulan dan lainnya terhadap 'Produksi'. Jika korelasinya positif, berarti ketika salah satu faktor naik, produksi juga cenderung naik. Jika korelasinya negatif, berarti salah satu faktor naik, produksi cenderung turun.

3) Data *Preparation* (Persiapan Data)

Lakukan transformasi data seperti normalisasi, standarisasi, dan mengubah data ke bentuk numerik.

```
1) bulan_angka = {'januari': 1, 'februari': 2, 'maret': 3, 'april': 4, 'mei': 5, 'juni': 6, 'juli': 7, 'agustus': 8, 'september': 9, 'oktober': 10, 'november': 11, 'desember': 12}
2) df['Bulan'] = df['Bulan'].map(bulan_angka)
3) df.head()
```

Pada Code mengubah data ke bentuk numerik dilakukan agar analisis data yang lebih mudah, kompatibilitas dengan analisis algoritma dan prediksi model, serta perhitungan statistik. Data numerik lebih mudah diinterpretasikan dalam konteks analisis dan mesin pembelajaran, meskipun perlu diingat bahwa konversi ini harus mempertimbangkan aspek kontekstual data sejak awal.

```
# Normalisasi
1. scaler = MinMaxScaler()
2. normalized_data = scaler.fit_transform(df)
3. normalized_data = pd.DataFrame(normalized_data, columns=df.columns)
4. print("Data yang dinormalisasi:\n", normalized_data.head())
# Standarisasi
5. scaler = StandardScaler()
6. standardized_data = scaler.fit_transform(df)
7. standardized_data = pd.DataFrame(standardized_data, columns=df.columns)
8. print("\nData yang distandarisasi:\n", standardized_data.head())
```

Pada Code dilakukan normalisasi data dan standarisasi data yang dijelaskan sebagai berikut :

Normalisasi (*MinMaxScaler*)

Pada bagian ini, data dari *DataFrame* *df* dinormalisasi menggunakan *MinMaxScaler*. Ini mengubah setiap nilai dalam setiap kolom menjadi rentang [0, 1] dengan mempertahankan proporsi relatif antara nilai-nilai tersebut. Hasil normalisasi disimpan dalam *DataFrame* baru *normalized_data*, dan rentang nilai-nilai ini dapat digunakan untuk membandingkan variabel dalam skala yang seragam.

Standarisasi (*StandardScaler*)

Pada bagian ini, data dari *DataFrame dfdistandarisasi* menggunakan *StandardScaler*. Ini menggeser dan mengubah setiap nilai dalam setiap kolom sehingga mereka memiliki rata-rata nol dan deviasi standar satu. Ini bermanfaat dalam mengatasi masalah variabel dengan skala yang berbeda dan juga ketika algoritma tertentu mengharapkan data dengan distribusi normal.

4) Model *Building* (Pembuatan Model)

Pilih metode regresi linier berganda untuk memprediksi hasil panen kelapa sawit, karena ini melibatkan beberapa variabel prediktor yang mempengaruhi hasil panen. Bagi data pelatihan menjadi fitur (variabel independen) dan label (variabel dependen), di mana fitur adalah variabel yang digunakan untuk memprediksi label. Latih model regresi linier berganda dengan data pelatihan.

```

1. # Feature Selection using RFE
2. X = normalized_data.drop('Produksi', axis=1)
3. y = normalized_data['Produksi']
4. # Create and train the linear regression model multiple_regression =
   LinearRegression()
5. rfe = RFE(multiple_regression, n_features_to_select=5) # Set the number of
   features you want to select
6. # Fit the RFE model to the data
7. rfe.fit(X, y)
8. # Get the selected features
9. selected_features = X.columns[rfe.support_]
10. # Print the selected features
11. print("Selected Features:")
12. print(selected_features)

```

Pada Code X dan y didefinisikan dimana X berisi semua kolom kecuali variabel target 'Produksi', dan y berisi variabel target. Model regresi linear (*multiple_regression*) dibuat. RFE (*Recursive Feature Elimination*) diterapkan pada model regresi linier. Parameter *n_features_to_select* menentukan berapa banyak fitur yang ingin Anda pilih. Model RFE dipasang ke data menggunakan *fit()* fungsi. *rfe.support_* mengembalikan topeng boolean yang menunjukkan fitur mana yang dipilih. Fitur yang dipilih diperoleh dengan menggunakan *X.columns[rfe.support_]*. Proses ini membantu untuk mengidentifikasi fitur yang paling relevan untuk memprediksi variabel target 'Produksi' menggunakan model regresi linier.

```

1) # Split data into training and testing sets
2) X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
   random_state=42)
3) # Create the linear regression model
4) multiple_regression = LinearRegression()

```

Dapat dilihat pada Code Fungsi *train_test_split* dari scikit-learn digunakan untuk membagi matriks fitur X dan vektor target y menjadi set pelatihan dan pengujian (*X_train, X_test, y_train, y_test*). Parameter *test_size* menentukan proporsi data yang harus dialokasikan untuk pengujian sebesar 20% data, dan *random_state* memastikan reproduktifitas pemisahan. Model regresi linier (*multiple_regression*) dibuat.

5) Model *Evaluation* (Evaluasi Model)

Evaluasi kinerja model menggunakan data pengujian yang tidak terlibat dalam pelatihan. Gunakan matrik evaluasi seperti *R-squared* dan *Root Mean Squared Error (RMSE)* untuk mengukur sejauh mana model dapat memprediksi hasil panen dengan akurat.

```

1. # Evaluate model performance on test set
2. mse = mean_squared_error(y_test, y_pred)
3. rmse = np.sqrt(mse)
4. r2 = r2_score(y_test, y_pred)
5. print("Root Mean Squared Error (RMSE):", rmse)
6. print("R2 Score (after feature selection):", r2)

```

Pada Code *mean_squared_error* from *scikit-learn* digunakan untuk menghitung mean squared error (MSE) antara nilai target aktual (y_{test}) dan nilai prediksi (y_{pred}). Akar kuadrat dari MSE dihitung menggunakan *np.sqrt()* *root mean squared error (RMSE)*. RMSE memberikan ukuran seberapa baik nilai prediksi selaras dengan nilai aktual dalam unit yang sama dengan variabel target. *r2_score* dari *scikit-learn* digunakan untuk menghitung koefisien determinasi (*R-squared*) antara nilai target aktual (y_{test}) dan nilai prediksi (y_{pred}). *R-squared* mengukur proporsi varian dalam variabel dependen yang dapat diprediksi dari variabel independen. Skor RMSE dan R-kuadrat dicetak untuk mengevaluasi performa model pada set pengujian. RMSE mengkuantifikasi kesalahan rata-rata antara nilai prediksi dan aktual, sedangkan *R-squared* menunjukkan proporsi varians yang dijelaskan oleh model.

Metode regresi linier dibentuk berdasarkan pola hubungan data signifikan masa lalu. Secara umum, prosedur pemecahan regresi linier dibagi sebagai 2 jenis yaitu regresi linier sederhana dan regresi linier berganda. Regresi linier sederhana adalah hubungan antara satu variabel terikat dengan satu variabel bebas, sedangkan regresi linier berganda adalah hubungan antara satu variabel terikat dengan dua atau lebih variabel bebas [9]. Variabel-variabel tersebut saling berhubungan atau memiliki hubungan sebab akibat. Metode regresi menjelaskan hubungan antara variabel-variabel tersebut [10]. Regresi linier berganda memperlihatkan hubungan variabel bebas dan variabel terikat dengan grafik berupa garis lurus. Perhitungan regresi linier memberikan persamaan yang dapat digunakan sebagai acuan untuk evaluasi nilai variabel terikat di masa mendatang dengan memasukkan nilai variabel bebas ke dalam persamaan [11]. Model persamaan regresi linier berganda yaitu:

$$y = a + b_1x_1 + \dots + b_nx_n \tag{1}$$

Keterangan:

- Y = variabel terikat (dependen)
- a = konstanta
- b1, bn = Koefisien regresi
- X1, Xn = Variabel bebas

Demi mendapatkan hasil ramalan yang akurat, perlu dilakukan evaluasi terhadap data yang sebenarnya. Beberapa metode biasanya digunakan untuk menghitung kesalahan prediksi salah satunya adalah *Root Mean Square Error (RMSE)* yang sering digunakan untuk mengevaluasi hasil peramalan. Ukuran perbedaan yang umum digunakan antara nilai yang diprediksi oleh model atau estimator dan nilai yang sebenarnya diamati adalah RMSE. Perbedaan individu disebut sebagai residual ketika dihitung menggunakan data sampel yang digunakan untuk estimasi, dan kesalahan prediksi ketika dihitung di luar sampel. RMSE memiliki fungsi yang menggabungkan besarnya kesalahan dalam prakiraan pada waktu yang berbeda menjadi satu ukuran kinerja prakiraan. RMSE adalah ukuran akurasi yang baik, tetapi hanya untuk membandingkan kesalahan perkiraan model yang berbeda untuk variabel tertentu, bukan antar variabel karena bergantung pada skala [10]. Mengenai perhitungan besarnya error yaitu perhitungan *root mean square error (RMSE)*. RMSE adalah akar dari MSE dan menggunakan metode berbasis gradien. Hasil prediksi dianggap baik jika semakin rendah nilai RMSE semakin baik prediksinya [12]. Rumus uji RMSE menurut persamaan adalah sebagai berikut.

$$RMSE = \sqrt{\frac{\sum (y_t - \hat{y}_t)^2}{n}} \tag{2}$$

Dimana:

- RMSE = Root Mean Square Error
- n = Jumlah Sampel
- y_t = Nilai Aktual Indeks
- \hat{y}_t = Nilai Prediksi Indeks

Model juga dapat dilihat dari sudut penerapannya; saat memeriksa penerapan model, persamaan kondisional (R^2) dapat diterapkan. Disebut koefisien determinasi karena 2 x 100% variasi variabel dependen Y dapat dijelaskan oleh variabel independen X yang diberi regresi linier Y pada X. Nilai koefisien determinasi kira-kira $0 < R^2 < 1$.

Artinya, bila R^2 mendekati 1, maka dapat dikatakan pengaruh variabel bebas terhadap variabel terikat sangat kuat. Berikut ini adalah interpretasi koefisien determinasi dapat dilihat dalam Tabel 1 [13].

Tabel 1 Interpretasi Koefisien Determinasi

Interval Koefisien	Tingkat Hubungan
0,80 - 1,00	Sangat Kuat
0,60 - 0,799	Kuat
0,40 - 0,599	Cukup Kuat
0,20 - 0,399	Rendah
0,00 - 0,199	Sangat Rendah

Rumus koefisien determinasi dapat dilihat pada persamaan (3) adalah sebagai berikut.

$$R^2 = 1 - \frac{SSR}{SST} \quad (3)$$

Di mana:

- SSR adalah *Sum of Squares Residual*, yaitu jumlah kuadrat dari selisih antara nilai aktual dan nilai prediksi ($y_t - \hat{y}_t$)
- SST adalah *Sum of Squares Total*, yaitu jumlah kuadrat dari selisih antara nilai aktual dan nilai rata-rata aktual

6) *Deployment* (Penyajian Hasil)

Jika model memenuhi kriteria kinerja yang memadai, implementasikan model dalam aplikasi atau sistem yang relevan untuk memprediksi hasil panen kelapa sawit.

```

1. # Buat prediksi pada data uji
2. y_pred = multiple_regression.predict(X_test)
3. # Buat DataFrame untuk membandingkan nilai 'Produksi' aktual dan nilai prediksi
4. predictions_df = pd.DataFrame({'Actual Produksi': y_test, 'Prediksi Produksi':
   y_pred})
5. # Tambahkan kolom selisih (error) antara nilai aktual dan prediksi
6. predictions_df['Error'] = predictions_df['Actual Produksi'] - predictions_df['Prediksi
   Produksi']
7. print(predictions_df)
8. # Plot nilai aktual vs prediksi
9. plt.figure(figsize=(10, 6))
10. # Plot titik-titik data aktual dengan warna merah
11. plt.scatter(y_test, y_test, color='red', label='Actual', alpha=0.5)
12. # Plot titik-titik data prediksi dengan warna biru
13. plt.scatter(y_test, y_pred, color='blue', label='Prediction', alpha=0.5)
14. plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
15. plt.xlabel('Actual Produksi')
16. plt.ylabel('Prediksi Produksi')
17. plt.title('Actual vs. Prediksi Produksi')
18. plt.show()

```

Dari Code hasil prediksi dan nilai aktual disusun dalam DataFrame yang memungkinkan perbandingan antara keduanya serta perhitungan selisih (*error*). Selanjutnya, *scatter plot* dibuat untuk memvisualisasikan perbandingan antara nilai aktual dan prediksi. Data aktual ditunjukkan dengan titik merah, prediksi data ditunjukkan dengan titik biru, dan garis putus-putus menggambarkan hubungan 1:1 antara nilai aktual dan prediksi. Plot ini memberikan gambaran tentang sejauh mana hasil prediksi model mendekati nilai aktual, serta perbedaan-perbedaan antara keduanya.

- 7) Perancangan sistem ini akan dilakukan menggunakan *Anaconda Navigator* dengan bahasa pemrograman *Python* dan statistik *library* untuk pengembangan model regresi

linier berganda. Library yang digunakan pada penelitian ini dapat dilihat pada Code dibawah

```

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. from sklearn.preprocessing import MinMaxScaler, StandardScaler
5. from sklearn.model_selection import train_test_split
6. from sklearn.linear_model import LinearRegression
7. from sklearn.metrics import mean_squared_error, r2_score
8. from sklearn.feature_selection import RFE
    
```

- *Numpy*(np alias): Pustaka yang kuat untuk perhitungan numerik dengan *Python*. Ini memberikan dukungan untuk *array*, matriks, fungsi matematika, dan operasi numerik lainnya.
- *Pandas*(pd alias): Pustaka populer untuk manipulasi dan analisis data. Ini menyediakan struktur data (seperti *DataFrame*) dan alat untuk menangani dan menganalisis data terstruktur secara efisien.
- *Matplotlib.pyplot*(plt alias): Pustaka untuk membuat visualisasi dan *plot* dengan *Python*. Ini menyediakan berbagai fungsi untuk membuat berbagai jenis *plot*.
- *MinMaxScaler* dan *StandardScaler* dari *sklearn.preprocessing*: Kelas-kelas ini digunakan untuk *preprocessing* data, khususnya untuk normalisasi atau standarisasi data untuk mempersiapkannya untuk algoritma pembelajaran mesin.
- *Train_test_split* dari *sklearn.model_selection*: Fungsi yang digunakan untuk membagi data menjadi set pelatihan dan pengujian. Ini penting untuk mengevaluasi kinerja model pembelajaran mesin pada data yang tidak terlihat.
- *LinearRegression* dari *sklearn.linear_model*: Kelas untuk membuat model regresi linier, yang biasa digunakan untuk analisis regresi dan tugas prediksi.
- *Mean_squared_error* dan *r2_score* dari *sklearn.metrics*: Fungsi-fungsi ini digunakan untuk mengevaluasi kinerja model regresi dengan menghitung metrik seperti *mean squared error (MSE)* dan *R-squared*.
- *RFE* dari *sklearn.feature_selection*: *Recursive Feature Elimination (RFE)* adalah teknik yang digunakan untuk memilih fitur yang paling relevan dari dataset, yang dapat meningkatkan performa model dan mengurangi *overfitting*.

4. Hasil dan Pembahasan

4.1 Implementasi sistem

Implementasi sistem adalah tahapan memenuhi kebutuhan pengguna dalam berinteraksi dengan sistem yang dibangun, berikut implementasi interface dari prediksi panen kelapa sawit.

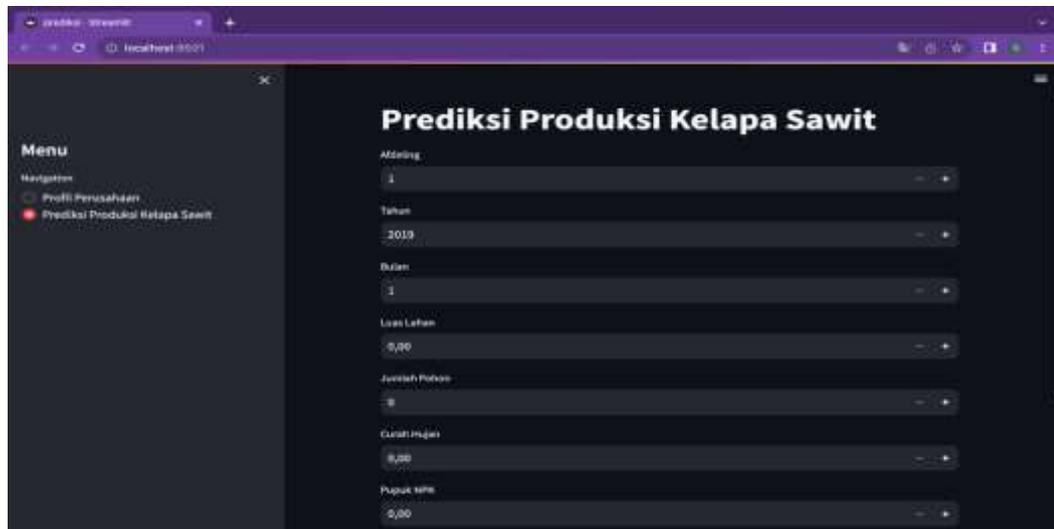
1) Tampilan Profil Perusahaan



Gambar 2 Tampilan Profil Perusahaan

Pada Gambar 2 merupakan tampilan saat pertama kali membuka aplikasi. Halaman ini penjelasan-penjelasan mengenai tentang Sejarah, Visi & Misi dan Lokasi.

2) Tampilan Prediksi



Gambar 3 Tampilan Prediksi

Gambar 3 adalah halaman yang akan tampil setelah admin berhasil login ke dalam aplikasi dengan memasukkan username dan password dengan benar. halaman ini juga tempat dilakukannya prediksi dengan menginput nilai adeling, tahun, bulan, luas lahan, jumlah pohon, curah hujan, pupuk NPK dan pupuk dolomite, yang kemudian dilakukan prediksi.

4.2 Hasil penelitian

Penelitian ini menggunakan regresi linier kelipatan untuk memprediksi hasil panen kelapa sawit di PTPN 13 Kebun Rimba Belian, Desa Semerangkai, Kecamatan Kapuas, Kabupaten Sanggau dengan metode CRISP-DM.

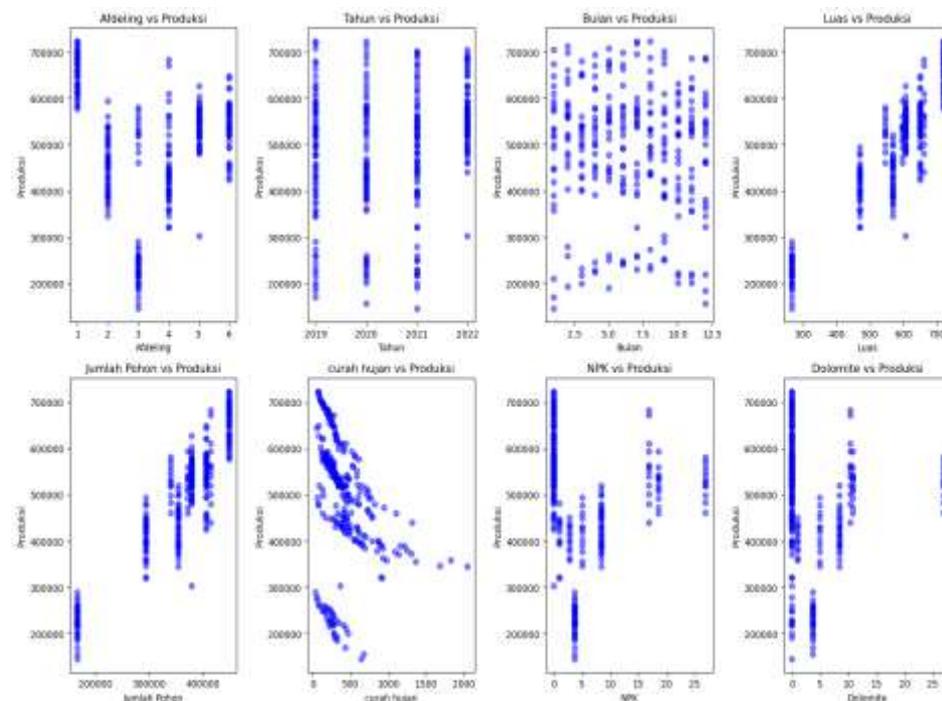
- 1) *Understanding the Business Problem* (Memahami Permasalahan Bisnis), tujuan dari penelitian ini yaitu memprediksi hasil panen kelapa sawit. Hasil prediksi akan digunakan untuk mendapatkan informasi gambaran dari prediksi hasil produksi kelapa sawit untuk masa yang akan datang, sehingga dapat dijadikan acuan dalam menentukan strategi untuk meningkatkan hasil produksi kelapa sawit. Tujuan lain bagi peneliti untuk lebih memahami metode regresi linier berganda untuk memprediksi hasil produksi kelapa sawit di PTPN 13 Kebun Rimba Belian. Data yang diambil data hasil produksi kelapa sawit tahun 2019-2022.
- 2) *Data Understanding* (Memahami Data), data yang relevan untuk prediksi hasil panen kelapa sawit, seperti data curah hujan, jumlah pohon, luas lahan, dan pupuk. Lakukan eksplorasi data untuk memahami karakteristik dan kualitas data, seperti melihat statistik deskriptif, visualisasi data, dan analisis korelasi.

Tabel 2 Deskripsi Data

	Bulan	Luas	Jumlah Pohon	curah hujan	NPK	Dolomite	Produksi
count	288	288	288	288	288	288	288
mean	6.5	567.1	354.457	3.967	4.47	3.2	490.164
std	3.4	135.1	84.439	2.935	6.95	6.0	133.014

	Bulan	Luas	Jumlah Pohon	curah hujan	NPK	Dolomite	Produksi
min	1	266.1	166.250	35	0	0	145.160
25%	3.75	525.3	328.437	212	0	0	419.272
50%	6.5	606.5	378.750	314	5	0	519.935
75%	9.250	649.5	406.250	481	5.85	4.02	576.972
max	12	719.5	449.375	2.049	26.90	26.90	723.220

Dari Tabel 2 dapat dilihat berapa jumlah data (*Count*) yang ada dalam dataset untuk setiap kolom. Nilai rata-rata (*Mean*) dari semua data dalam suatu kolom. Ini diperoleh dengan menjumlahkan semua data dan kemudian membaginya dengan jumlah total data. *Standar Deviasi (Std)* ukuran sebaran data sekitar rata-rata. *Standar deviasi* yang lebih tinggi menunjukkan bahwa data lebih tersebar dari rata-rata, sedangkan standar deviasi yang lebih rendah menunjukkan data lebih mendekati rata-rata. Nilai terkecil (*Min*) dalam suatu kolom. Ini adalah batas bawah dari rentang data. Nilai Kuartil Pertama yang membagi data menjadi dua bagian 25% data lebih rendah darinya dan 75% data lebih tinggi darinya. Ini memberikan gambaran tentang bagaimana data yang tersebar di bagian bawah rentang. Nilai tengah dalam dataset saat data diurutkan. Ini adalah titik di mana 50% data lebih rendah dan 50% data lebih tinggi. Nilai yang membagi data menjadi dua bagian: 25% data lebih tinggi darinya dan 75% data lebih rendah darinya. Ini memberikan gambaran tentang bagaimana data yang tersebar di bagian atas rentang. Nilai tertinggi (*Maks*) dalam suatu kolom. Ini adalah batas atas dari rentang data.



Gambar 4 Visualisasi Data

Dari Gambar 4 dapat diketahui hubungan antara fitur dan label, pola pada luas dan jumlah pohon menunjukkan bahwa semakin luas lahan dan jumlah pohon akan menghasilkan produksi lebih banyak. Pada pola curah hujan, NPK, dan

dolomite dapat disimpulkan bahwa semakin rendah nilainya, hasil produksi semakin tinggi. Akan tetapi terjadi penyimpangan pada NPK dan dolomite yang seharusnya semakin tinggi nilai akan semakin banyak nilai produksi. Hal ini dapat terjadi dikarenakan ada pengaruh dari jumlah pohon dan luas pohon.

```
1. Bulan      -0.077589
2. Luas       0.915159
3. Jumlah Pohon 0.914846
4. curah hujan -0.319553
5. NPK        -0.103530
6. Dolomite   -0.071942
7. Produksi   1.000000
Name: Produksi, dtype: float64
```

Dari kode dapat disimpulkan nilai korelasi antara produksi dengan variabel lain dalam dataset menunjukkan bahwa luas dan jumlah pohon memiliki hubungan positif yang kuat dengan produksi, sementara tahun juga memiliki hubungan positif yang sedang. Sementara itu, curah hujan, NPK, dan dolomite memiliki hubungan negatif yang lemah dengan produksi. Variabel afdeling dan bulan memiliki hubungan yang lemah dan tidak konsisten dengan produksi.

- 3) Data *Preparation* (Persiapan Data), pembersihan data untuk mengatasi masalah seperti data yang hilang, duplikat, atau tidak valid. Lakukan transformasi data seperti normalisasi, standarisasi, dan mengubah data ke bentuk numerik.

Tabel 3 Data yang telah distandarisasi

Bulan	Luas	Jumlah Pohon	curah hujan	NPK	Dolomit	Produksi
-1,593255	1.129	1.126	0,543	-0,644	-0,536516	0,713835
-1,303572	1.129	1.126	-0,364	-0,644	-0,536516	1,074271
-1,01389	1.129	1.126	-0,749	-0,644	-0,536516	1.436.363
-0,724207	1.129	1.126	-0,470	-0,644	-0,536516	1.190.249
-0,434524	1.129	1.126	-1,033	-0,644	-0,536516	1.635.785

Tabel 3 menunjukkan data yang telah distandarisasi. Nilai-nilai dalam setiap kolom telah diubah sedemikian rupa sehingga memiliki rata-rata 0 dan standar deviasi 1. Hal ini memungkinkan perbandingan yang lebih adil antara variabel-variabel dengan skala yang berbeda dalam statistik analisis.

Tabel 4 Data yang telah dinormalisasi

Bulan	Luas	Jumlah Pohon	curah hujan	NPK	Dolomit	Produksi
0	1.0	1.0	0,258689	0	0	0,760803
0,090909	1.0	1.0	0,126614	0	0	0,843598
0,181818	1.0	1.0	0,070506	0	0	0,926772
0,272727	1.0	1.0	0,111221	0	0	0,870238
0,363636	1.0	1.0	0,029295	0	0	0,972581

Dalam Tabel 4, nilai-nilai dalam setiap kolom telah diubah sedemikian rupa sehingga berada dalam kisaran 0 hingga 1, dengan mempertahankan proporsi relatif antara nilai-nilai tersebut. Tujuan dari normalisasi adalah membuat data memiliki skala yang seragam dan dapat diterapkan dengan baik dalam beberapa algoritma machine learning dan analisis statistik.

- 4) Model *Building* (Pembuatan Model), metode regresi linier berganda dipilih untuk memprediksi hasil panen kelapa sawit, Bagi data pelatihan menjadi fitur (variabel independen) dan label (variabel dependen) jumlah data yang dibagi yaitu 20% dari total data akan digunakan untuk data testing, sedangkan 80% akan digunakan untuk data training yang berarti jumlah data pelatihan 230 dan data pengujian 58, dimana semua fitur digunakan untuk memprediksi label.
- 5) Model *Evaluation* (Evaluasi Model), *R-squared* dan *Root Mean Squared Error (RMSE)* untuk mengukur sejauh mana model dapat memprediksi hasil panen dengan akurat.

Root Mean Squared Error (RMSE): 0.06978246915965547
 R2 Score : 0.9306036013117357

Analisis evaluasi model regresi linier pada Code diatas menunjukkan bahwa model memiliki performa yang baik dalam memprediksi nilai target. *Root Mean Squared Error (RMSE)* dengan nilai 0.0698 mengindikasikan bahwa model memiliki rata-rata error prediksi sekitar 0.069, sementara *R-squared (R2) Score* sebesar 0.9306 menandakan bahwa model mampu menjelaskan variasi data sekitar 93.06%. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi nilai target dan menjelaskan variasi data dengan baik. Hasil perhitungan manual menggunakan 5 data aktual dan prediksi pada tabel 5.

Tabel 5 Data Aktual dan Data Prediksi

No	Actual Produksi(yt)	Prediksi Produksi(ŷt)	(yt-ŷt)	(yt-ŷt) ²
1	0,705324707	0,674546399	0,030778308	0,000947
2	0,511780784	0,693032468	-0,181251684	0,032852
3	0,605957859	0,649062049	-0,04310419	0,001858
4	0,525187697	0,468209412	0,056978284	0,003247
5	0,495069716	0,465387781	0,029681935	0,000881
Total	2,843321	2,950238	-0,10692	0,039785

Menghitung manual RMSE dapat dilihat dibawah ini:

$$\begin{aligned}
 RMSE &= \sqrt{\frac{\sum (yt - \hat{y}t)^2}{n}} \\
 &= \sqrt{\frac{0,039785}{5}} \\
 &= 0,089202
 \end{aligned}$$

Koefisien determinasi, atau *R-squared (R2)*, mengukur seberapa baik variabilitas dalam variabel target (yt) dapat dijelaskan oleh model Anda (ŷt). R2 berkisar antara 0 hingga 1, di nilai mana yang lebih tinggi menunjukkan bahwa model lebih baik dalam menjelaskan data variabilitas. Anda dapat menghitung R2 menggunakan rumus berikut:

$$R^2 = 1 - \frac{SSR}{SST}$$

Cari nilai SSR

$$SSR = \sum_{i=1}^n ((y_t - \hat{y}_t)^2 = 0,039785$$

Nilai SSR didapat 0,039785, selanjutnya cari nilai SST

$$SST = \sum_{i=1}^n (y_t - \bar{y})^2$$

Dimana nilai \bar{y} adalah nilai rata-rata dari nilai aktual (y_t) dan n banyaknya data sehingga di dapatkan rumus :

$$\bar{y} = \frac{\sum_{i=1}^n y_t}{n}$$

$$\bar{y} = \frac{2,843321}{5} = 0,5686642$$

Kemudian kita dapat menghitung SST :

$$SST = \sum_{i=1}^n (y_t - \bar{y})^2$$

$$SST = (0,705324707 - 0,5686642)^2 + (0,511780784 - 0,5686642)^2 + (0,605957859 - 0,5686642)^2 + (0,525187697 - 0,5686642)^2$$

$$SST = 0.01442102069 + 0.00384173335 + 0.00143032741 + 0.00386833046 + 0.00416654745 = 0.02772895936$$

Didapat nilai SSR = 0,039785 dan SST = 0.02772895936

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{0,039785}{0,02772895936} = 0,433102913$$

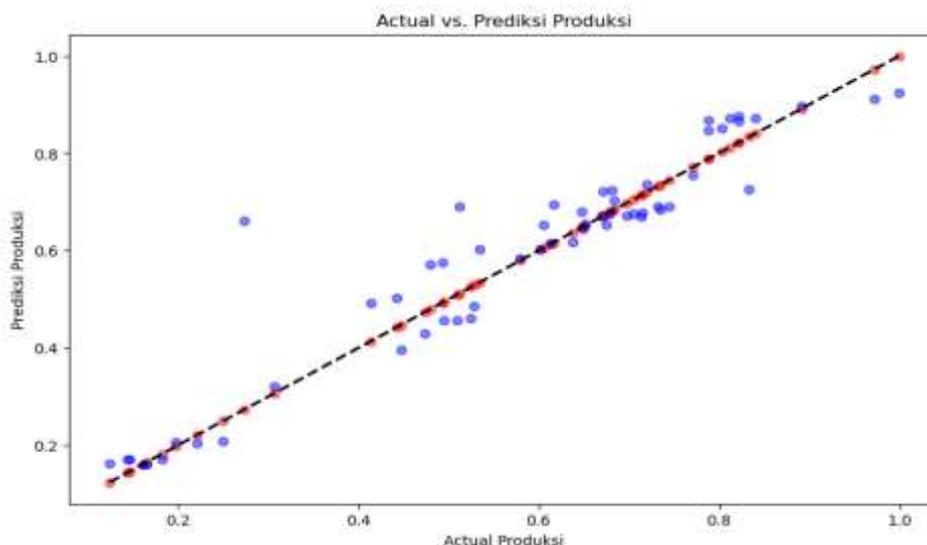
Jadi hasil perhitungan manual menggunakan sampel data aktual dan hasil prediksi yaitu nilai RMSE yang dihitung sebesar 0,089202, Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi nilai target dan menjelaskan variasi data dengan baik sedangkan nilai *R-Squared* (R^2) sekitar 0,433102913, yang menunjukan bahwa sekitar 43,31% variabilitas dalam variabel target dapat dijelaskan.

- 6) *Deployment* (Penyajian Hasil), model memenuhi kriteria kinerja yang memadai. Pada fase ini dijelaskan tahap cara menyimpan dan menggunakan model yang disimpan, pertama adalah menyimpan pemodelan `ipynb` yang dibangun kedalam file `joblib` menggunakan `library` Python dengan code seperti pada Code dibawah.

```
#save model
1. import joblib
2. joblib.dump(multiple_regression, 'regression_model.joblib')
```

Setelah berhasil menyimpan pemodelan dalam file `joblib` langkah selanjutnya yaitu dengan memanggil file `joblib` tersebut kedalam aplikasi `website` yang dibangun menggunakan `library Python` yaitu `Streamlit` kedalam file `py` seperti code seperti code dibawah.

```
1. import streamlit as st
2. import joblib
3. #Load model
4. loaded_model = joblib.load('regression_model.joblib')
```



Gambar 5 Plot Hasil Prediksi dan Data Aktual

Dari Gambar 5 merupakan plot hasil prediksi dan data aktual. Data aktual ditunjukkan pada warna merah dan hasil prediksi ditunjukkan pada warna biru. Dari hasil tersebut dapat disimpulkan bahwa hasil prediksi memiliki kecenderungan mendekati garis linier. Hal ini membuktikan bahwa akurasi yang dimiliki cukup baik. Sedangkan memiliki performa yang baik dalam memprediksi nilai target. *Root Mean Squared Error (RMSE)* dengan nilai 0.0698 mengindikasikan bahwa model memiliki rata-rata error prediksi sekitar 0.069, sementara *R-squared (R2) Score* sebesar 0.9306 menandakan bahwa model mampu menjelaskan variasi data sekitar 93.06%. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi nilai target dan menjelaskan variasi data dengan baik.

5. Simpulan

Kesimpulan dari penelitian ini adalah penulis dapat menerapkan *data mining* untuk memprediksi hasil panen kelapa sawit di PTPN 13 Kebun Rimba Belian menggunakan metode regresi linier berganda dengan hasil *Root Mean Squared Error (RMSE)* sebesar 0.0698 mengindikasikan bahwa model memiliki rata-rata error prediksi sekitar 0.069, sementara *R-squared (R2) Score* sebesar 0.9306 menandakan bahwa model mampu menjelaskan variasi data sekitar 93.06%. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi nilai target dan menjelaskan variasi data dengan baik. Dilakukan perhitungan manual menggunakan 5 data aktual dan prediksi dengan hasil RMSE yang dihitung sebesar 0.089202 dan nilai *R-Squared (R2)* sekitar , yang menunjukkan bahwa sekitar 43.31%.

Meskipun metode regresi linier berganda memberikan hasil yang baik dalam penelitian ini, disarankan untuk mempertimbangkan menguji menggunakan metode regresi lainnya, sehingga dapat membantu membandingkan kinerja berbagai metode dan memperoleh pemahaman yang lebih mendalam tentang mana metode yang paling cocok untuk kasus ini. Melakukan validasi tambahan terhadap model yang telah dibangun dengan menggunakan dataset yang berbeda atau teknik *cross-validation*. Ini akan memberikan gambaran yang lebih akurat tentang sejauh mana model dapat digeneralisasi ke data yang belum pernah dilihat sebelumnya. Peningkatan data aktual dengan memperoleh lebih banyak data aktual panen kelapa sawit dari PTPN 13 Kebun Rimba Belian dapat membantu memperbaiki kualitas dan akurasi model. Semakin banyak data yang digunakan untuk melatih model, semakin baik model dapat memahami pola dan tren yang ada.

Daftar Referensi

[1] Sulardi, *Budidaya Tanaman Kelapa Sawit Buku Ajar*, Pertama. PT Dewangga Energi Internasional Anggota IKAPI (403/JBA/2021), 2022.
 [2] A. A.-F. Nur Wahyudin, A. Primajaya, and A. S. Y. Irawan, "Penerapan Algoritma Regresi Linear Berganda Pada Estimasi Penjualan Mobil Astra Isuzu," *Techno.Com*, vol. 19, no.

- 4, pp. 364–374, 2020, doi: 10.33633/tc.v19i4.3834.
- [3] E. Rahayu, I. Parlina, and Z. A. Siregar, “Application of Multiple Linear Regression Algorithm for Motorcycle Sales Estimation,” *Jomlai J. Mach. Learn. Artif. Intell.*, vol. 1, no. 1, pp. 1–10, 2022, doi: 10.55123/jomlai.v1i1.142.
- [4] N. Rusmilawati and P. T. Prasetyaningrum, “Implementation of Data Mining in the Prediction of Palm Oil Production for PT Borneo Ketapang Indah Using the Linear Regression Method,” *J. Inf. Syst. Artif. Intell.*, vol. 1, no. 2, pp. 1–7, 2021.
- [5] M. Adha, E. Utami, and H. Hanafi, “Prediksi Produksi Jagung Menggunakan Algoritma Apriori Dan Regresi Linear Berganda (Studi Kasus : Dinas Pertanian Kabupaten Dompu),” *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 3, pp. 803–820, 2022, doi: 10.29100/jipi.v7i3.3139.
- [6] I. Solihat, J. Jasmi, and F. Fachruddin, “Prediksi Jumlah Produksi Kelapa Sawit Dengan Regresi,” *Snastikom*, vol. 1, no. 1, pp. 17–26, 2020.
- [7] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, “Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 364, 2019, doi: 10.22146/jnteti.v8i4.537.
- [8] Y. Suhanda, I. Kurniati, & S. Norma, “Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik,” *Jurnal Teknologi Informatika dan Komputer*, vol. 6, no. 2, pp. 12-20, 2020.
- [9] A. N. Maharadja, I. Maulana, and B. A. Dermawan, “Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi,” *J. Appl. Informatics Comput.*, vol. 5, no. 1, pp. 95–102, 2021, doi: 10.30871/jaic.v5i1.3184.
- [10] R. Puspasari, S. Effendi, H. Kurniawan, M. Ayoe, and E. Nasution, “Penentuan Prediksi Hasil Panen Kelapa Sawit Menggunakan Metode Regresi Linier,” *Pros. Semin. Nas. Ris. Dan Inf. Sci.*, vol. 4, pp. 91–98, 2022.
- [11] T. Indarwati, T. Irawati, and E. Rimawati, “Penggunaan Metode Linear Regression Untuk Prediksi Penjualan Smartphone,” *J. Teknol. Inf. dan Komun.*, vol. 6, no. 2, pp. 2–7, 2019, doi: 10.30646/tikomsin.v6i2.369.
- [12] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
- [13] A. Pranata, M. Akbar Hsb, T. Akhdansyah, and S. Anwar, “Penerapan Metode Pemulusan Eksponensial Ganda dan Tripel Untuk Meramalkan Kunjungan Wisatawan Mancanegara ke Indonesia,” *J. Data Anal.*, vol. 1, no. 1, pp. 32–41, 2018, doi: 10.24815/jda.v1i1.11873.