Contents list avaliable at Directory of Open Access Journals (DOAJ)

JUTIN : Jurnal Teknik Industri Terintegrasi

Volume 7 Issue 1 2024, Page 262-272

ISSN: 2620-8962 (Online)

Journal Homepage: https://journal.universitaspahlawan.ac.id/index.php/jutin/index

Penerapan Data Mining Produksi Padi di Pulau Sumatera Menggunakan Analisis Regresi Linear

Yohanes Nababan[1]✉, Isna Nugraha[2]

Teknik Industri Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

DOI: 10.31004/jutin.v7i1.23545

✉ Corresponding author:
[20032010140@student.upnjatim.ac.id]

Article Info

Kata kunci:
Analisis regresi linear
CRISP-DM
Poduksi Padi
Phyton

Abstrak

Indonesia, yang secara utama merupakan negara agraris, sangat mengandalkan pertanian sebagai mata pencaharian, terutama dalam produksi padi. Padi merupakan komoditas penting, khususnya di Sumatera. Memahami faktor-faktor berpengaruh seperti curah hujan, kelembapan, suhu rata-rata, dan luas panen sangat penting untuk produksi padi yang efektif. Penelitian ini menerapkan metode CRISP-DM: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, dan Pemodelan. Analisis regresi linier berganda digunakan dengan menggunakan bahasa pemrograman Python di Google Colab untuk menilai dampak faktor-faktor ini terhadap produksi padi. Hasil menunjukkan bahwa curah hujan, kelembapan, dan suhu rata-rata tidak signifikan memengaruhi produksi padi, sementara luas panen memengaruhi secara signifikan. Model regresi diungkapkan sebagai Y = 12,3X1 + 1637,1X2 – 159677,3X3 + 5,1X4. Model ini memberikan wawasan berharga bagi para petani untuk memprioritaskan faktor-faktor berpengaruh dalam produksi padi di masa depan.

Keywords:
Linear Regression Analysis
CRISP-DM
Rice Production
Phyton

Abstract

Indonesia, primarily an agrarian nation, relies heavily on farming as a livelihood, particularly in rice production. Rice is a crucial commodity, especially in Sumatra. Understanding the influential factors such as rainfall, humidity, average temperature, and harvest area is vital for effective rice production. This research applies the CRISP-DM method: Business Understanding, Data Understanding, Data Preparation, and Modeling. Multiple linear regression analysis is employed using Python programming in Google Colab to assess the impact of these factors on rice production. Results indicate that rainfall, humidity, and average temperature insignificantly affect rice production, while harvest area significantly influences it. The regression model is expressed as Y = 12.3X1 + 1637.1X2 – 159677.3X3 + 5.1X4. This model provides valuable insights for farmers to prioritize influential factors in future rice production.

Received 28 December 2023; Received in revised form 31 December 2023 year; Accepted 6 January 2024
Available online 17 January 2024 /

1. INTRODUCTION

Indonesia sebagai negara agraris yang dikenal memiliki berbagai jenis sumber daya alam yang melimpah. Salah satu pulau di Indonesia merupakan penghasil tanaman pangan terbesar. Tanaman pangan yang dimaksudkan disini adalah padi. (Dyah Pitaloka, 2022). Padi merupakan tanaman pangan penting karena menghasilkan beras yang menjadi sumber bahan makanan pokok. Padi merupakan komoditas utama di Indonesia dalam menyokong pangan masyarakat (Dungu et al., 2023). Ada beberapa alasan penting untuk meningkatkan produksi beras secara berkelanjutan. Faktor produksi memiliki hubungan yang sangat erat dengan produk akhir dalam proses pembuatannya (Alamri et al., 2022). Faktor-faktor penting yang mempengaruhi produktivitas padi dan memberikan dasar bagi pengembangan strategi yang lebih efektif dalam meningkatkan produktivitas padi di Pulau Sumatera seperti curah hujan, suhu, dan kelembapan memiliki pengaruh yang signifikan terhadap pertumbuhan dan produktivitas padi. Untuk itu menganalisis penggunaan faktor- faktor yang mempengaruhi tingkat produksi padi mengunakan teknik data *mining*. (Mayasari et al., 2023).

Data mining adalah metode untuk menemukan pola tertentu dari kumpulan data yang berjumlah besar (Prastiwi et al., 2022). Seiring berjalannya waktu, suatu data akan terus bertambah dari data yang sekarang digabungkan dengan data di masa depan sehingga akan ada aliran data yang besar, algoritma data mining berfungsi secara efektif dan efisien untuk menganalisis data yang besar (Sekar Setyaningtyas et al., 2022). Data mining merupakan proses iterative dan interaktif untuk menemukan pola atau model baru yang sempurna, bermanfaat dan dapat dimengerti dalam suatu database yang sangat besar (*massive database*) (Sikumbang, 2018). Data mining menggambarkan sebuah pengumpulan teknik-teknik dengan tujuan untuk menemukan pola-pola yang tidak diketahui pada data yang telah dikumpulkan. Data mining merupan suatu alat yang memungkinkan para pengguna untuk mengakses secara cepat data dengan jumlah yang besar (Ahmad et al., 2022). Data mining digunakan untuk mendapatkan pengetahuan / informasi yang baru. Prinsip kerjanya melakukan pemisahan suatu data yang besar menjadi data yang lebih kecil atau intisarinya dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data yang cukup besar (Srirahayu & Pribadie, 2023).

Data Mining mengidentifikasi fakta-fakta atau kesimpulan-kesimpulan yang disarankan berdasarkan penyaringan melalui data untuk menjelajahi pola-pola atau anomali-anomali data. Data Mining mempunyai 5 fungsi: *classification*, yaitu menyimpulkan definisi- definisi karakteristik sebuah grup. *Clustering*, yaitu mengindentifikasikan kelompok-kelompok dari barang-barang atau produk-produk yang mempunyai karakteristik khusus. *Association*, yaitu mengidentifikasikan hubungan antara kejadian-kejadian yang terjadi pada suatu waktu, seperti isi-isi dari keranjang belanja. *Sequencing*, hampir sama dengan *association*, *sequencing* mengidentifikasikan hubungan-hubungan yang berbeda pada suatu periode waktu tertentu, seperti pelanggan-pelanggan yang mengunjungi supermarket secara berulang-ulang. *Forecasting* memperkirakan nilai pada masa yang akan datang berdasarkan pola-pola dengan sekumpulan data yang besar, seperti peramalan permintaan pasar (Ahmad et al., 2022).

Analisis regresi merupakan salah satu alat analisis satistika yang memanfaatkan hubungan antara dua variabel atau lebih. Tujuannya adalah untuk membuat perkiraan (prediksi) yang dapat dipercaya untuk nilai suatu variabel (biasa disebut variabel terikat atau variabel dependen atau variabel respons), jika nilai variabel lain yang berhubungan dengannya diketahui (biasa disebut variabel bebas atau variabel independen atau variabel prediktor) (Purba & Purba, 2022). Analisis Regresi adalah suatu statistik yang memanfaatkan hubungan statistik antar dua atau lebih variabel kuantitatif sehingga satu variabel dapat diprediksi dari variabel yang lainnya . Atau lebih sering disebut variabel dependen dan variabel independen. Regresi Linear Berganda (RLB) merupakan pengembangan dari Regresi Linear Sederhana (RLB). Pada analisis RLB, variabel dependen dipengaruhi oleh lebih dari satu variabel independen (Herdiana, 2022).

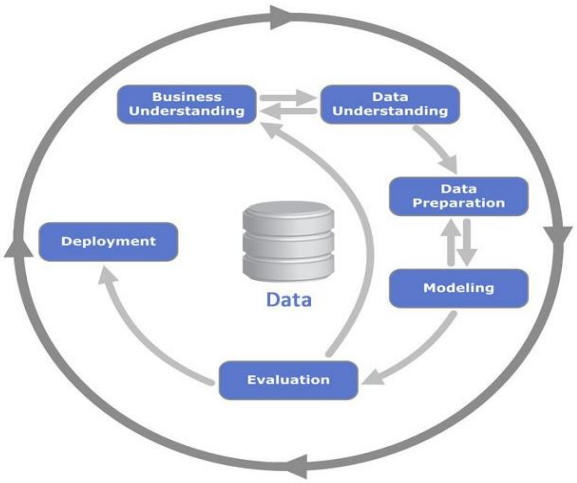
Model regresi linier berganda (*multiple linear regression*) mengasumsikan sebuah hubungan linier (dalam parameter) antara sebuah hubungan linier (dalam parameter) antara sebuah variabel respon y (*dependent variable*) dan sebuah himpunan dari variabel prediktor (Puteri & Silvanie, 2020). Analisis ini digunakan untuk melihat sejumlah variabel independen X1 , X2 , … Xk terhadap variabel dependen Y berdasarkan nilai variabel-variabel independen X1 , X2 , … Xk. Dalam regresi berganda seluruh variabel bebas dimasukkan kedalam perhitungan regresi serentak. Dengan demikian diperoleh persamaan regresi guna memprediksi variabel terikat dengan memasukkan secara serentak serangkaian variabel bebas. Dalam persamaan regresi dihasilkan konstanta dan koefisien regresi bagi masing-masing variabel bebas (Wisudaningsi et al., 2019). Melalui perhitungan regresi

linear akan menghasilkan persamaan yang dapat dijadikan acuan untuk memperkirakan nilai variabel dependent di waktu mendatang dengan memasukan nilai variabel independent ke dalam persamaan (Maharadja et al., 2021).

CRISP-DM adalah metode yang menyediakan proses standar dalam data mining untuk memecahkan masalah dalam bisnis. CRISP-DM lebih mudah diterapkan karena setiap tahapan atau fase didefinisikan dan terstruktur dengan jelas serta memiliki metodologi data mining yang lengkap dan terdokumentasi dengan baik (Yudiana et al., 2023). Pada penelitian ini menggunakan metode analisis regresi linear untuk meramalkan produktivitas padi terhadap faktor-faktor yang mempengaruhi pertumbuhan padi di indonesia khususnya di pulau Sumatera seperti kelembapan, curah hujan, suhu-rata-rata dan luas panen. Perbandingan data yang diperoleh adalah data mulai dari tahun 2010 sampai 2020. Tujuan penelitian ini untuk megetahui pengaruh faktor-faktor yang mempengaruhi produktivitas padi. Dengan demikian para petani dapat melihat apa saja yang mempengaruhi produktivitas padi dan mempertimbangkan kedepannya.

2. METHODS

Penelitian ini mengadopsi pendekatan *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai metodenya. CRISP-DM adalah metode yang menyediakan proses standar dalam data mining untuk memecahkan masalah dalam bisnis. CRISP-DM lebih mudah diterapkan karena setiap tahapan atau fase didefinisikan dan terstruktur dengan jelas serta memiliki metodologi data mining yang lengkap dan terdokumentasi dengan baik. Sejak tahun 1996 standar CRISP-DM sudah dikembangkan guna menjadikan proses industri bisnis dalam sebuah penelitian. William Vorheis, salah satu pencetus CRISP-DM (dari Data Science Central). Terdapat 6 proses tahapan dalam metode CRISP DM yaitu, *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* dan *Deployment* (Yudiana et al., 2023). Proses Metode CRISP-DM dapat dilihat pada Gambar 2.1.



Gambar 2.1 Proses CRISP-DM

Sumber : Kenneth Jensen

2.1 Pemahaman Bisnis (*Business Understanding Phase*)

Pemahaman masalah pada penelitian ini adalah produksi padi di Indonesia, khususnya provinsi yang ada di pulau Sumatera. Pada tahapan ini diperlukan pemahaman tentang apa saja faktor-faktor yang mempengaruhi produktivitas padi. Tujuan dari pemahaman ini untuk memprediksi hasil produksi padi berdasarkan faktor-faktor yang dapat mempengaruhi laju panen padi.

2.2 Pemahaman Data (*Data Understanding Phase*)

Pada pemahaman data berdasarkan data produksi padi khususnya berbagai provinsi di pulau Sumatera didapatkan data berdasarkan tahun 2010 sampai 2020 dalam bentuk CSV yang didapatkan dari situs https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera. Kemudian untuk deskripsi dataset tanaman padi berbentuk tabel yang berisi 88 baris dan 7 kolom yang terdiri dari nama provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata. Kemudian untuk atribut yang dipilih adalah produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata.

2.3 Persiapan Data (*Data Preparation*)

Penelitian diawali dengan pengumpulan data dari situs *Kaggle*. Peneliti melakukan seleksi data, memilih informasi yang relevan untuk diolah. Data yang dipilih kemudian disaring dan dipindahkan ke *Microsoft Excel* untuk diintegrasikan menjadi satu tabel. Setelah itu, data diubah ke format CSV untuk mempermudah proses berikutnya. Langkah selanjutnya melibatkan pengolahan data mentah untuk memastikan bahwa data yang dipilih memenuhi syarat untuk proses analisis. Kegiatan ini mencakup pembersihan data dari duplikasi, penghapusan informasi yang tidak diperlukan, serta pengecekan keberadaan *missing value* pada dataset. Kemudian *checking* data yang dilakukan dengan perintah menampilkan informasi dataset dan deskripsi data apakah sudah sesuai dengan data yang dada di dalam csv.

```
# Tampilkan informasi dataset
print(data.info())

# Statistik deskriptif
print(data.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88 entries, 0 to 87
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Provinsi         88 non-null     object
 1   Tahun            88 non-null     int64
 2   Produksi         88 non-null     float64
 3   Luas Panen       88 non-null     float64
 4   Curah hujan      88 non-null     float64
 5   Kelembapan       88 non-null     float64
 6   Suhu rata-rata   88 non-null     float64
dtypes: float64(5), int64(1), object(1)
memory usage: 4.9+ KB
None
             Tahun       Produksi     Luas Panen  Curah hujan  Kelembapan  \
count     88.0000  8.800000e+01      88.000000    88.000000   88.000000
mean    2015.0000  1.962717e+06  366335.745341  2497.511705   81.142841
std        3.1804  1.362043e+06  244096.564858   999.183978    4.417833
min     2010.0000  2.308740e+05   63142.040000   327.330000   54.200000
25%     2012.0000  5.903730e+05  141418.500000  1756.650000   79.000000
50%     2015.0000  2.016610e+06  364209.000000  2297.600000   82.000000
75%     2018.0000  3.021008e+06  546691.977500  3178.750000   83.855000
max     2020.0000  4.881089e+06  872737.000000  5228.000000   87.730000

       Suhu rata-rata
count       88.000000
mean        26.616477
std          0.915481
min         22.190000
25%         26.357500
50%         26.800000
75%         27.100000
max         28.800000
```

Gambar 2. Tampilan Informasi dan Deskripsi Dataset

```
#Mencari dan Menangani missing values
print(data.isnull().sum())
```

```
Provinsi          0
Tahun             0
Produksi          0
Luas Panen        0
Curah hujan       0
Kelembapan        0
Suhu rata-rata    0
dtype: int64
```

Gambar 3. Hasil Pencarian Data Kosong

2.4 Pemodelan (*Modelling*)

Pada tahap ini, setelah data dikumpulkan dan disiapkan, langkah selanjutnya adalah melakukan pengolahan data menggunakan metode analisis regresi linear. Proses ini akan memanfaatkan platform Google Colab dengan menggunakan bahasa pemrograman Python. Untuk menjalankan analisis data, beberapa langkah perlu diikuti:

1) Olah data dan menentukan *datasheet*
2) Memilih *library* yang diperlukan
3) Mempersiapkan google colab dan memindahkan data dari google drive
4) Menentukan variabel x dan y
5) Menentukan data *training* dan data *test*
6) Korelasi antar data
7) Menghitung hasil regresi linear

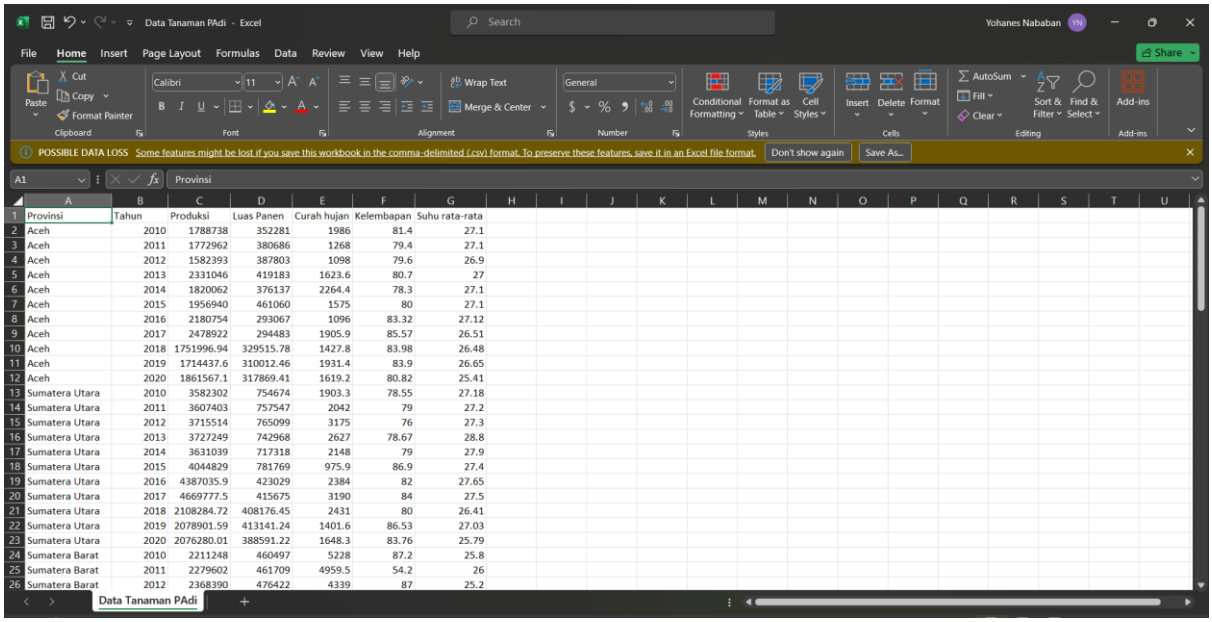
8) Membandingkan data asli dan data prediksi

3. RESULT AND DISCUSSION

Berdasarkan implementasi metode analisis regresi lienar menggunakan pemrograman python, berikut adalah hasil analisisnya :

1) Olah data dan menentukan *datasheet*

Proses pengolahan data yang digunakan disimpan dalam bentuk file CSV. Data yang diolah sebanyak 88 baris dan 7 kolom.



Provinsi	Tahun	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
Aceh	2010	1788738	352281	1986	81.4	27.1
Aceh	2011	1772962	380686	1268	79.4	27.1
Aceh	2012	1582393	387803	1098	79.6	26.9
Aceh	2013	2331046	419183	1623.6	80.7	27
Aceh	2014	1820062	376137	2264.4	78.3	27.1
Aceh	2015	1956940	461060	1575	80	27.1
Aceh	2016	2180754	293067	1096	83.32	27.12
Aceh	2017	2478922	294483	1905.9	85.57	26.51
Aceh	2018	1751996.94	329515.78	1427.8	83.98	26.48
Aceh	2019	1714437.6	310012.46	1931.4	83.9	26.65
Aceh	2020	1861567.1	317869.41	1619.2	80.82	25.41
Sumatera Utara	2010	3582302	754674	1903.3	78.55	27.18
Sumatera Utara	2011	3607403	757547	2042	79	27.2
Sumatera Utara	2012	3715514	765099	3175	76	27.3
Sumatera Utara	2013	3727249	742968	2627	78.67	28.8
Sumatera Utara	2014	3631039	717318	2148	79	27.9
Sumatera Utara	2015	4044829	781769	975.9	86.9	27.4
Sumatera Utara	2016	4387035.9	423029	2384	82	27.65
Sumatera Utara	2017	4669777.5	415675	3190	84	27.5
Sumatera Utara	2018	2108284.72	408176.45	2431	80	26.41
Sumatera Utara	2019	2078901.59	413141.24	1401.6	86.53	27.03
Sumatera Utara	2020	2076280.01	388591.22	1648.3	83.76	25.79
Sumatera Barat	2010	2211248	460497	5228	87.2	25.8
Sumatera Barat	2011	2279602	461709	4959.5	54.2	26
Sumatera Barat	2012	2368390	476422	4339	87	25.2

Gambar 4. Data CSV

Pada gambar 3.1 memberikan data jumlah produksi padi pada 5 provinsi di Pulau Sumatera dalam kurun waktu 2010-2020 yaitu Aceh, Sumatera Utara, Riau, Sumatera Selatan, Lampung, Bengkulu yang berbebntuk CSV.

2) Memilih *library* yang diperlukan

Proses awal dalam melakukan perhitungan analisis regresi berganda dengan menggunakan bahasa pemrograman python adalah menentukan library yang digunakan dalam proses analisis dan memindahkan data ke drive. Langkah-langkah nya adalah

```
pip install pandas matplotlib scikit-learn
```

```
[90] import numpy as np
     import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn import metrics
     import matplotlib.pyplot as plt
```

Gambar 5. Pemilihan Library

Import Numpy adalah perintah dalam bahasa pemrograman Python yang digunakan untuk mengimpor (menggunakan) pustaka NumPy (Numerical Python) yang berguna untuk operasi numerik dan matematika di Python. Pandas adalah pustaka Python yang sangat populer untuk analisis data dan manipulasi data.

3) Mempersiapkan *google colab* dan memindahkan data dari *google drive*

Proses ini diperlukan karena penggunaan google colab harus terkoneksi dengan internet dan data yang ada harus disimpan di drive *google*. *File* data2.csv yang disimpan di komputer lokal akan diunggah di drive dengan nama *file* yang sama.

```
[3] data = pd.read_csv("/content/drive/MyDrive/MSIB/Data Tanaman PAdi.csv")
```

Gambar 6. Menghubungkan Google Colap dengan Drive

Gambar 3.3 proses penghubungan google colap dengan data yang telah disimpan dalam bentuk CSV di dalam goggle drive.

Hasil proses *running script* di atas ditampilkan sebagai berikut.

```
[4]  print(data.head())

   Provinsi  Tahun   Produksi  Luas Panen  Curah hujan  Kelembapan  \
0      Aceh   2010  1788738.0    352281.0       1986.0        81.4
1      Aceh   2011  1772962.0    380686.0       1268.0        79.4
2      Aceh   2012  1582393.0    387803.0       1098.0        79.6
3      Aceh   2013  2331046.0    419183.0       1623.6        80.7
4      Aceh   2014  1820062.0    376137.0       2264.4        78.3

   Suhu rata-rata
0            27.1
1            27.1
2            26.9
3            27.0
4            27.1
```

Gambar 7. Hasil *Running Script*

Gambar 3.4 menjelaskan secara keseluruhan digunakan untuk mencetak beberapa baris pertama dari DataFrame ke konsol agar Anda dapat dengan cepat memeriksa dan memahami struktur dan konten dari DataFrame tersebut. Ini sangat berguna saat Anda bekerja dengan data besar dan ingin mendapatkan gambaran singkat tentang bagaimana data tersebut terlihat atau sebelum melakukan manipulasi lebih lanjut.

4) Menentukan variabel x dan y

Proses awal adalah membedakan data yang menjadi variabel dependen (y) dan variabel indepeneden (x). Kolom yang menjadi variabel dependent (y) adalah kolom produksi dan yang menjadi variabel independent (x) adalah kolom luas panen, curah hujan, kelembapan, dan suhu rata-rata.

```
x = data [['Curah hujan', 'Kelembapan', 'Suhu rata-rata', 'Luas Panen']]
y = data ['Produksi']
```

Gambar 8. Pemilihan variabel x dan y

Pada gambar 3.5 menejelaskan proses pemilihan variabel dependen (y) dan variabel independen (x).

5) Menentukan data *training* dan data *test*

Data sekitar 88 data yang diolah, data tersebut dibagi 80% menjadi data training dan 20% digunakan untuk data test. Data training digunakan untuk menghitung prediksi regresi berganda dan data test digunakan untuk perhitungkan data prediksi dan dapat dibandingkan antara data asli dengan data hasil prediksi.

```
# Pisahkan data menjadi set pelatihan dan pengujian
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Gambar 9. Penentuan Data *Training* dan *Test*

Persentase data yang akan dialokasikan untuk set pengujian. Dalam hal ini, test size = 0,2 berarti 20% dari data akan diambil untuk set pengujian, sementara 80% akan digunakan untuk set pelatihan.

6) Korelasi antar data

Analisis korelasi digunakan untuk melihat keterkaitan tentang derajat hubungan variabel sehingga dapat mengetahui hubungan variabel yang ada. Korelasi positif menunjukkan bahwa ketika satu variabel naik, yang lain cenderung naik juga. Sebaliknya, korelasi negatif menunjukkan bahwa ketika satu variabel naik, yang lain cenderung turun.

```
# Menghitung korelasi antar variabel
correlation_matrix = data.corr()

# Tampilkan matriks korelasi
print(correlation_matrix)
```

```
                                                Suhu rata-rata
                    Tahun  Produksi  Luas Panen     -0.286212  embapan
Tahun            1.000000 -0.106727   -0.255376      0.097289  .118833
Produksi        -0.106727  1.000000    0.890992      0.199955  .064575
Luas Panen      -0.255376  0.890992    1.000000     -0.352095  .054906
Curah hujan     -0.176658 -0.033895   -0.085239      0.164621  .007745
Kelembapan       0.118833 -0.064575   -0.054906      1.000000  .000000
Suhu rata-rata -0.286212  0.097289    0.199955     -0.352095  0.164621
```

Gambar 10. Korelasi Antar Data

Pada gambar 3.4 menjelaskan tentang hubungan anatara setiap variabel. Dimana korelasi antara luas panen dengan curah hujan bernilai negatif. Kemudian untuk korelasi antara luas penen dan kelembapan bernilai negatif. Kemudian untuk korelasi antara luas panen dan suhu rata-rata bernilai positif.

7) Menghitung hasil regresi linear

Langkah selanjutnya adalah menghitung regresi linear. Library yang digunakan adalah library sklean.

```
# Mencari nilai intercept
intercept_value = model.params[0]

print("Nilai Intercept:", intercept_value)
```

```
Nilai Intercept: 4166987.3499577483
```

Gambar 11. proses penghitungan nilai koefisien

```
# Mencari nilai koefisien
coefficients = model.coef_

print("Nilai Koefisien:", coefficients)
```

```
Nilai Koefisien: [ 1.23197856e+01  1.63718249e+03 -1.59677261e+05  5.08524520e+00]
```

Gambar 12. Lanjutan proses penghitungan nilai koefisien

Pada Gambar 3.8 menjelaskan nilai koefisien tiap-tiap kolom didapakanlah untuk kolom curah hujan sebesar 12.3197856, untuk kelembapan sebesar 1637.18249, untuk suhu rata-rata sebesar −159677.261, dan untuk luas tanah sebesar 5.0852452. Berdasarkan hasil pada perhitungan di atas, nilai regresi linearnya adalah : $Y = 12{,}3X_1 + 1637{,}1X_2 - 159677{,}3X_3 + 5{,}1X_4$

Evaluasi untuk model linear regresi

Evaluasi untuk model linear regresi dapat dilakukan dengan beberapa cara evaluasi seperti menggunakan *mean absolute error* (MAE), *mean squared error* (MSE) atau *root mean squared error* (RMSE)

```
# Prediksi pada set pengujian
y_pred = model.predict(x_test)

# Evaluasi model
r_squared = metrics.r2_score(y_test, y_pred)
mse = metrics.mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

print(f'R-squared: {r_squared}')
print(f'Mean Squared Error: {mse}')
print("Root Mean Squared Error (RMSE):", rmse)
```

```
R-squared: 0.6185405972542619
Mean Squared Error: 631835773038.161
Root Mean Squared Error (RMSE): 794880.9804229568
```

Gambar 13. Hasil perhitungan MSE dan RMSE

Pada gambar 3.9 menjelaskan nilai dari hasil perhitungan nilai r-square sebesar 0,6185. Kemudian nilai MSE (*Mean Square Error*) sebesar 631835773038,161. Dan untuk nilai RMSE (*Root Mean Squared Error*) sebesar 794880,98.

8) Membandingkan data asli dan data prediksi

Pada tahap ini akan dilakukan perbandingan antara nilai y awal dengan nilai y prediksi. Nilai y prediksi dihitung pada rumus regresi linear berganda yang telah dihitung pada langkah di atas.

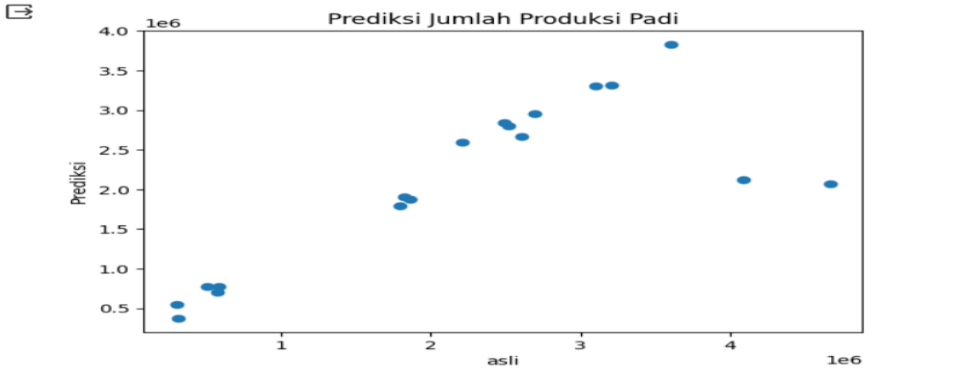
```
# Misalnya, y_test adalah nilai sebenarnya dan y_pred adalah hasil prediksi
data = {'Data Asli': y_test, 'Hasil Prediksi': y_pred}
df = pd.DataFrame(data)

# Menampilkan tabel
print(df)
```

```
     Data Asli  Hasil Prediksi
76    296925.16    5.494032e+05
0    1788738.00    1.788903e+06
26   2519020.00    2.796869e+06
22   2211248.00    2.596224e+06
12   3607403.00    3.830573e+06
67    502552.00    7.755445e+05
10   1861567.10    1.878297e+06
18   4669777.50    2.066495e+06
4    1820062.00    1.908571e+06
68    581910.00    7.718254e+05
85   2488641.91    2.840143e+06
65   2696877.46    2.957738e+06
53    309932.68    3.740699e+05
80   3207002.00    3.311326e+06
84   4090654.00    2.123108e+06
64   2603396.24    2.662220e+06
33    574864.00    7.041904e+05
79   3101455.00    3.302157e+06
```

Gambar 14. hasil perhitungan membandingkan data asli dengan data prediksi

Hasil proses pada gambar 3.10 dapat dibuat dalam bentuk grafik. Gambar 3.8 Hasil y asli dengan y prediksi dalam bentuk grafik. Hasil grafik memperlihatkan antara data asli dan data prediksi menghasilkan selisih angka antara data aktual dan data prediksi yang lumayan besar.

```
# Visualisasi prediksi vs. aktual
plt.scatter(y_test, y_pred)
plt.xlabel('asli')
plt.ylabel('Prediksi')
plt.title('Prediksi Jumlah Produksi Padi')
plt.show()
```



Gambar 15. Grafik y asli dengan y prediksi

Gambar 3.11 menjelaskan tentang model visualization antar data aktual dan data prediksi pada jumlah produksi padi.

Perhitungan dengan library statsmodels

Library lain yang dapat digunakan untuk melakukan analisis regresi linear adalah dengan menggunakan library statsmidel. Hasil perhitungan dengan library statsmodel ada pada gambar 3.9.

```
import statsmodels.api as sm
x = x_train.to_numpy()
y = y_train.to_numpy()
x = sm. add_constant(x)
model = sm.OLS(y,x).fit ();
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.840
Model:                            OLS   Adj. R-squared:                  0.830
Method:                 Least Squares   F-statistic:                     85.32
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           3.85e-25
Time:                        06:31:05   Log-Likelihood:                -1024.1
No. Observations:                  70   AIC:                             2058.
Df Residuals:                      65   BIC:                             2069.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4.167e+06   2.31e+06      1.802      0.076    -4.5e+05    8.78e+06
x1             12.3198     73.688      0.167      0.868    -134.845     159.484
x2           1637.1825   1.49e+04      0.110      0.913   -2.81e+04    3.13e+04
x3          -1.597e+05   8.11e+04     -1.969      0.053   -3.22e+05    2242.733
x4              5.0852      0.279     18.222      0.000       4.528       5.643
==============================================================================
Omnibus:                       52.355   Durbin-Watson:                   2.235
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              177.683
Skew:                           2.421   Prob(JB):                     2.61e-39
Kurtosis:                       9.122   Cond. No.                     1.50e+07
==============================================================================
```

Gambar 16. Hasil perhitungan dengan library Statsmodel

Analisa :

Berdasarkan gambar 3.9 dapat dilihat keseluruhan hasil dari metode analisis regresi linear . Dimana nilai dari R-square sebesar 0,84 kemudian nilai dari adj R-square sebesar 0,83. Kemudian untuk nilai p-value dari koefisien x1 sebesar 0,868 kemudian koefisien x2 0,013 kemudian koefisien x3 sebesar 0,053 dan koefisien x4 sebesar 0. Kemudian didapakanlah untuk kolom curah hujan sebesar 12.3197856, untuk kelembapan sebesar 1637.18249, untuk suhu rata-rata sebesar −159677.261, dan untuk luas tanah sebesar 5.0852452. Dari hasil tersebut untuk kedepannya para petani lebih mengerti apa saja yang paling menonjol untuk mempengaruhi dari produksi padi tanpa harus menghiraukan faktor-faktor lain.

4. CONCLUSION

Dalam penelitian ini menentukan pengaruh faktor-faktor seperti curah hujan, kelembapan, suhu rata-rata ,dan luas panen dalam produksi padi di berbagai provinsi di Pulau Sumatera. Metode yang digunakan pada penelitian ini adalah analisis regresi linear berganda dimana penelitian ini menggunakam variabel dependennya sebagai produksi padi dan untuk variabel independennya sebagai curah hujan, kelembapan, suhu rata-rata dan luas panen. Untuk nilai p-*value* dari masing-masing variabel dimana variabel x1 sebagai curah hujan memiliki nilai sebesar 0,868; untuk variabel x2 sebagai kelembapan memiliki nilai sebesar 0,913; untuk variabel x3 sebagai suhu rata-rata memiliki nialai sebesar 0,053 dimana ketiga variabel ini tidak terdapat pengaruh yang signifikan terhadap produksi padi. Kemudian pada variabel x4 sebagai luas panen memiliki nilai p-*value* sebesar 0 dimana pada variabel ini terdapat pengaruh yang signifikan terhadap produksi padi. Untuk nilai koefisien tiap-tiap variabel didapakanlah untuk variabel curah hujan sebesar 12.3197856, untuk variabel kelembapan sebesar 1637.18249, untuk variabel suhu rata-rata sebesar −159677.261, dan untuk variabel luas tanah sebesar 5.0852452. Berdasarkan hasil pada perhitungan di atas, nilai regresi linearnya adalah : $Y = 12{,}3X_1 + 1637{,}1X_2 – 159677{,}3X_3 + 5{,}1X_4$. Hal ini membuktikan bahwa dalam penggunaan google colap dengan bahasa pemrogaman phyton sangat berguna dalam melakukan analisis menggunakan metode analisis regresi berganda karena sangat mudah memahami perhitungan dalam penelitian. Kelemahan pada penelitian ini kurangnya data pada masa sekarang dikarenakan data yang diberikan yaitu sampai pada tahun 2020. Pada penelitian mendatang, dapat memperluas dataset produksi padi dengan melibatkan faktor yang lain yang mempengaruhi produksi padi. dan mempertimbangkan penggunaan metode analisis regresi linear untuk membuat penelitian lebih komprehensif. Hal ini akan memungkinkan perbandingan yang lebih mendalam serta memberikan kontribusi pada peningkatan kualitas dan kelengkapan penelitian.

5. ACKNOWLEDGMENTS (Optional)

In this section, you can acknowledge any support given, which is not covered by the author's contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

6. REFERENCES

Ahmad, I., Samsugi, S., & Irawan, Y. (2022). Implementasi Data Mining Sebagai Pengolahan Data. *Jurnal Teknoinfo*, *16*(1), 46. http://portaldata.org/index.php/portaldata/article/view/107

Alamri, M. H., Rauf, A., & Saleh, Y. (2022). Analisis Faktor-Faktor Produksi Terhadap Produksi Padi Sawah Di Kecamatan Bintauna Kabupaten Bolaang Mongondow Utara. *AGRINESIA: Jurnal Ilmiah Agribisnis*, *6*(3), 240–249. https://doi.org/10.37046/agr.v6i3.16145

Dungu, A. R., Umbu, E., & Retang, K. (2023). Faktor-Faktor Yang Mempengaruhi Produksi Padi Sawah Tadah Hujan Di Desa Umbu Pabal Kecamatan Umbu Ratu Nggay Barat Kabupaten Sumba Tengah Factors Affecting Rice Production of Rainfed Rice in the Village of Umbu Pabal, Umbu Ratu Nggay Barat District, Centr. *Jurnal Pertanian Agros*, *25*(1), 714–723.

Dyah Pitaloka, S. (2022). Analisis Faktor Produksi Padi Di Jawa Timur Tahun 2005-2015 Dengan Metode Cobb-Douglass. *Growth: Jurnal Ilmiah Ekonomi Pembangunan*, *1*(2), 93–100.

Herdiana, A. (2022). Studi Kasus Kemiskinan Di Indonesia Level Provinsi Dan Faktor-Faktor Yang Mempengaruhinya Menggunakan Regresi Linear Berganda. *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *10*(1), 89–93. https://doi.org/10.24252/msa.v10i1.23361

Maharadja, A. N., Maulana, I., & Dermawan, B. A. (2021). Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi. *Journal of Applied Informatics and Computing*, *5*(1), 95–102. https://doi.org/10.30871/jaic.v5i1.3184

Mayasari, R., Nugraha, B., Juwita, A. R., & Heryana, N. (2023). Analisis Produktifitas Padi di Pulau Sumatera menggunakan Exploratory Data Analysis (EDA). *Jurnal Elektronik Sistem Informasi Unsika*, *1*(1), 17–24.

Prastiwi, H., Jeny Pricilia, & Errissya Rasywir. (2022). Implementasi Data Mining Untuk Menentuksn Persediaan Stok Barang Di Mini Market Menggunakan Metode K-Means Clustering. *Jurnal Informatika Dan Rekayasa Komputer(JAKAKOM)*, *2*(1), 141–148. https://doi.org/10.33998/jakakom.2022.2.1.34

Purba, D., & Purba, M. (2022). Aplikasi Analisis Korelasi dan Regresi menggunakan Pearson Product Moment dan Simple Linear Regression. *Citra Sains Teknologi*, *1*(2), 97–103.

Puteri, K., & Silvanie, A. (2020). Machine Learning untuk Model Prediksi Harga Sembako. *Jurnal Nasional Informatika*, *1*(2), 82–94.

Sekar Setyaningtyas, Indarmawan Nugroho, B., & Arif, Z. (2022). Tinjauan Pustaka Sistematis: Penerapan Data Mining Teknik Clustering Algoritma K-Means. *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, *10*(2), 52–61. https://doi.org/10.21063/jtif.2022.v10.2.52-61

Sikumbang, E. D. (2018). Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori. *Jurnal Teknik Komputer AMIK BSI (JTK)*, *Vol 4*, *No.*(September), 1–4.

Srirahayu, A., & Pribadie, L. S. (2023). Review Paper Data Mining Klasifikasi Data Mining. *Jurnal Ilmiah Informatika Global*, *14*(April). http://ejournal.uigm.ac.id/index.php/IG/article/view/2981%0Ahttp://ejournal.uigm.ac.id/index.php/IG/article/download/2981/1841

Wisudaningsi, B. A., Arofah, I., & Belang, K. A. (2019). Pengaruh Kualitas Pelayanan Dan Kualitas Produk Terhadap Kepuasan Konsumen Dengan Menggunakan Metode Analisis Regresi Linear Berganda. *Statmat : Jurnal Statistika Dan Matematika*, *1*(1), 103–117. https://doi.org/10.32493/sm.v1i1.2377

Yudiana, Y., Yulia Agustina, A., & Nur Khofifah, dan. (2023). Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan. *Indonesian Journal of Islamic Economics and Business*, *8*(1), 01–20. http://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb