PERBANDINGAN METODE REGRESI LINEAR DAN K-NEAREST NEIGHBOR (KNN) DALAM MEMPREDIKSI PRODUKSI TANAMAN PADI DI PULAU SUMATERA

Kiki Mustaqim[1)*], Noviana Riza[2)], Yusuf[3)], Muhammad Reefy Hidayatullah[4)]
[1,2,3,4)] Universitas Logistik dan Bisnis Internasional, Fakultas Logistik, Teknologi dan Bisnis, Program Studi Sains Data
kiki@ulbi.ac.id[1)], novianariza@ulbi.ac.id[2)], yusufnasihin@gmail.com[3)], muhreefy16@gmail.com[4)]

Received: 30 December 2024 Accepted: 12 January 2025 Published: 20 January 2025

*kiki@ulbi.ac.id

Kata Kunci: Produksi Padi, Prediksi, KNN, Regresi Linear

Abstrak : Padi merupakan bahan pangan yang sangat penting untuk menunjang kebutuhan pangan di Indonesia, khususnya di Pulau Sumatera. Faktor-faktor yang memengaruhi produksi padi meliputi luas panen, kelembapan, curah hujan, dan suhu rata-rata. Setiap tahun, suhu bumi yang terus meningkat akibat pemanasan global berdampak pada iklim yang fluktuatif, sehingga dapat menghambat produksi padi. Memahami faktor-faktor tersebut menjadi penting untuk pengembangan strategi yang efektif dalam meningkatkan produktivitas padi. Penelitian ini menggunakan bahasa pemrograman Python pada Google Colab untuk membandingkan metode regresi linear berganda dan *K-Nearest Neighbors* (KNN) dalam memprediksi produksi padi di Pulau Sumatera. Hasil penelitian menunjukkan bahwa metode regresi linear lebih akurat dibandingkan KNN, dengan nilai R^2 regresi linear sebesar 0,868181. Selain itu, regresi linear memiliki nilai MAE yang lebih rendah yaitu 324967,05 dan nilai MSE yang lebih rendah yaitu 1,571551e+11. Sedangkan, hasil penelitian menggunakan metode KNN menunjukkan nilai R^2 sebesar 0,703748, nilai MAE 416812,35 dan nilai MSE 3,531935e+11. Hasil ini menunjukkan bahwa regresi linear lebih andal dalam memprediksi produksi padi di Pulau Sumatera dan dapat digunakan sebagai alat bantu dalam pengambilan keputusan strategis di sektor pertanian.

PENDAHULUAN

Indonesia sebagai negara agraris dengan sumber daya alam yang sangat melimpah. Banyaknya sumber daya alam tersebut menjadi salah satu alasan mengapa Indonesia banyak diminati oleh negara lain. Salah satu tanaman pangan terbesar yang dihasilkan Indonesia adalah padi [1]. Padi merupakan bahan pangan yang sangat penting untuk menunjang nilai pangan yang ada di Indonesia. Produksi padi menduduki peringkat ketiga setelah gandum dan diikuti oleh jagung. Data tersebut berlaku di Benua Asia yang mana menjadi tempat bagi para petani yang telah memproduksi lebih dari 90% dari total produksi beras dunia [2]. Produksi padi dipengaruhi oleh beberapa faktor, diantaranya adalah luas panen, kelembapan, suhu rata-rata, dan curah hujan. Indonesia beriklim tropis dengan curah hujan yang tinggi. Namun, perubahan iklim dalam beberapa tahun terakhir membuat curah hujan tidak menentu, mengakibatkan produksi padi menjadi terganggu dan tidak stabil. Untuk mengatasi kondisi cuaca yang tidak menentu, diperlukan metode untuk memprediksi produksi padi berdasarkan data yang ada. Salah satu metode yang dapat digunakan adalah regresi linear berganda dan *K-Nearest Neighbor* (KNN).

Regresi linear adalah metode untuk menganalisis hubungan antara variabel bebas (X) dan variabel terikat (Y) [3]. Dalam memprediksi produksi padi pada penelitian ini, terdapat 5 faktor yang digunakan sebagai variabel bebas (X), yaitu luas panen, curah hujan, kelembapan, dan suhu rata-rata, sedangkan produksi padi menjadi variabel terikat (Y). *K-Nearest Neighbor* (KNN) merupakan sebuah metode algoritma yang biasanya digunakan dalam mengklasifikasi objek dari data yang jaraknya paling dekat dengan objek yang akan diklasifikasi [4].

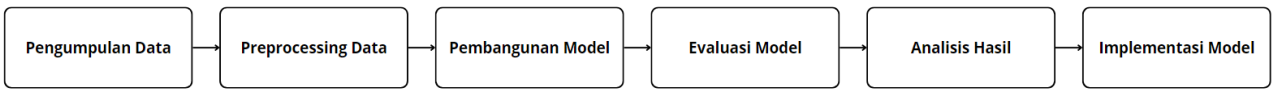
Berdasarkan survei Kerangka Sampel Area (KSA) oleh Badan Pusat Statistik (BPS), Pulau Sumatera menjadi penghasil padi nasional terbesar ketiga setelah Pulau Jawa dan Sulawesi Selatan. Data BPS 2024 mencatat produksi padi di Pulau Sumatera pada 2023 meningkat 83.970 ton atau 6,11% dibandingkan tahun 2022 yang sebesar 1.373 ton [5]. Regresi linear dan *K-Nearest Neighbor* (KNN) dipilih sebagai metode utama dalam penelitian ini karena keunggulannya dalam menangani data numerik dan hubungan antar variabel yang kompleks. Regresi linear efektif untuk menganalisis hubungan linier antara variabel bebas dan terikat, sementara KNN dapat menangani pola tidak linier tanpa membutuhkan asumsi distribusi data tertentu. Meskipun metode lain seperti *decision tree* dan *random forest* dipertimbangkan, regresi linear dan KNN lebih relevan karena kesederhanaannya, kemudahan interpretasi, dan keandalan dalam memprediksi produksi padi berdasarkan faktor-faktor yang ada [6]. Hasil penelitian ini dapat membantu pemerintah dan instansi terkait dalam merencanakan strategi produksi padi yang lebih efektif, mengelola risiko akibat perubahan iklim, mengalokasikan sumber daya dengan tepat, mendukung kebijakan ketahanan pangan, serta mendorong modernisasi sektor pertanian melalui adopsi teknologi berbasis data. Hal ini berkontribusi langsung pada peningkatan produktivitas dan kesejahteraan petani, khususnya di Pulau Sumatera.

TINJAUAN LITERATUR

Yohanes Lababan dkk (2024) pada [1] melakukan sebuah penelitian yang berjudul "Penerapan *Data Mining* Produksi Padi di Pulau Sumatera Menggunakan Analisis Regresi Linear" menunjukkan nilai R^2 pada model regresi linear berganda untuk memprediksi produksi padi di Pulau Sumatera bernilai 0,618. Maka dibutuhkan metode lain yang lebih akurat dalam memprediksi produksi padi. Berdasarkan beberapa penelitian, metode KNN merupakan salah satu metode yang efektif dalam memprediksi. Namun, belum ada penelitian yang menggunakan metode KNN dalam memprediksi produksi padi di Pulau Sumatera. Diajeng Sekar Seruni dkk (2020) pada [7] melakukan penelitian untuk memprediksi pertumbuhan jumlah penduduk di Kota Malang menggunakan metode KNN Regression. Hasil pengujian menunjukkan nilai Mean Absolute Percentage Error (MAPE) yang diperoleh adalah 0.02526%. Sedangkan rata-rata nilai MAPE untuk prediksi hingga 24 bulan ke depan adalah 0.13506%. Hal ini menunjukkan bahwa metode KNN Regression cukup akurat dalam memprediksi pertumbuhan penduduk di Kota Malang dalam beberapa tahun ke depan. Ervan Triyanto, dkk (2019) pada [8] Melakukan penelitian untuk memprediksi produksi padi di Kabupaten Bantul menggunakan Regresi Linear, hasil yang diperoleh adalah nilai mean absolute deviation (MAD) adalah 0,101 dengan data pelatihan dari tahun 2009 - 2017. Persamaan regresi yang didapatkan adalah Y = 8307,561443282 + 5,9294543706657x1 + 118,28063200866x1 + 175,71009241484x3. Penelitian ini berkontribusi pada literatur dengan membandingkan metode regresi linear dan K-Nearest Neighbor (KNN) dalam memprediksi produksi padi di Pulau Sumatera. Berbeda dari penelitian sebelumnya, studi ini tidak hanya mengevaluasi performa kedua metode secara empiris tetapi juga menyoroti kelebihan masing-masing dalam menangani karakteristik data lokal yang kompleks. Dengan demikian, penelitian ini memberikan wawasan baru tentang penerapan metode prediksi yang lebih sesuai untuk wilayah spesifik.

METODE PENELITIAN

Penelitian ini bertujuan membandingkan metode regresi linear dan KNN dalam memprediksi produksi tanaman padi di Pulau Sumatera. Adapun alur penelitiannya adalah sebagai berikut.



Gambar 1 Alur Penelitian

Alur pada penelitian ini terdiri dari pengumpulan data produksi padi, preprocessing data, pembangunan model, evaluasi model, analisis hasil, dan implementasi model. Setiap proses memiliki tujuan tersendiri untuk memastikan hasil dapat tercapai.

1. Pengumpulan Data

Pengumpulan data merupakan langkah awal untuk memahami suatu kejadian atau masalah dengan mengekstrak informasi atau data dari beberapa sumber [9]. Pada penelitian ini, data diperoleh dari situs Kaggle dengan tambahan data dari situs BPS dan Badan Meteorologi, Klimatologi dan Geofisika (BMKG). Penelitian ini menggunakan data produksi padi, luas panen, curah hujan, kelembapan, dan suhu rata-rata dari tahun 2012 hingga 2020, meskipun dataset tersedia sejak 1993. Pemilihan data mulai dari tahun 2012 dilakukan untuk memastikan relevansi terhadap kondisi pertanian dan lingkungan terkini, mengingat perubahan iklim, teknologi pertanian, dan kebijakan pemerintah yang berkembang pesat dalam dekade terakhir. Selain itu, data dari rentang waktu ini dianggap lebih representatif untuk mencerminkan tren dan pola yang relevan dengan situasi saat ini, sehingga hasil prediksi menjadi lebih akurat dan bermanfaat bagi pengambilan keputusan.

2. *Preprocessing* Data
   Proses *preprocessing* sangat penting dilakukan sebelum penerapan model untuk memastikan kualitas data dan meningkatkan akurasi prediksi yang dihasilkan [8]. Tahapan preprocessing mencakup penggabungan data dari berbagai sumber, seperti Kaggle, BPS, dan BMKG, untuk memperoleh informasi yang lebih lengkap dan memastikan integrasi data yang tepat. Selanjutnya, dilakukan pemeriksaan nilai kosong, penghapusan data duplikat, analisis statistik deskriptif, pemeriksaan persebaran data, pemeriksaan *outlier*, matriks korelasi, pemilihan fitur, serta *scaling* data untuk mempersiapkan dataset agar siap digunakan dalam analisis dan prediksi.

3. Train Test Split
   Train test split adalah proses pembagian dataset menjadi data latih (*train*) dan data uji (*test*). Data latih digunakan untuk membangun dan melatih model. Setelah model dibangun, model diuji akurasinya menggunakan data uji [10]. Pada penelitian ini, data latih yang digunakan adalah 80% dari dataset, sedangkan data uji yang digunakan adalah 20% dari dataset.

4. Penerapan Regresi Linear
   Regresi linear merupakan proses membangun model prediksi dengan menjelaskan hubungan antara variabel independen (X) dan variabel dependen (Y). Terdapat dua jenis regresi linear, yaitu regresi linear sederhana dan regresi linear berganda. Regresi linear sederhana adalah sebuah metode analisis statistik yang membahas tentang hubungan suatu variabel bebas dengan variabel terikat, sedangkan regresi linear berganda membahas hubungan lebih dari satu variabel bebas dengan variabel terikat [3]. Pada penelitian ini, metode yang digunakan adalah regresi linear berganda dikarenakan pada dataset yang digunakan, terdapat satu variabel terikat (Y) dan 4 variabel bebas (X). Variabel terikat tersebut adalah atribut Produksi, sedangkan variabel bebasnya adalah Luas panen, Curah hujan, Kelembapan, dan Suhu rata-rata.

5. Penerapan *K-Nearest Neighbor* (KNN)
   K-Nearest Neighbor (KNN) adalah sebuah metode prediksi menggunakan algoritma dengan nilai hasil ditentukan dari nilai titik-titik terdekat dari nilai yang ingin diprediksi [7]. Langkah-langkah yang harus diikuti dalam analisis adalah sebagai berikut.
   a. Menentukan jumlah k atau jumlah data terdekat.
   b. Menghitung jarak atribut pada dataset dengan atribut pada data yang ingin diprediksi. Satuan jarak yang biasa digunakan adalah jarak Euclidean.
   c. Data diurutkan berdasarkan jarak yang terdekat dari nilai yang ingin diprediksi
   d. Nilai pada variabel target dihitung dengan mengambil rata-rata dari sejumlah data terdekat sesuai dengan jumlah k yang telah ditentukan.

6. Perbandingan Metode

Metode yang diteliti dibandingkan berdasarkan nilai koefisien determinasi (R^2), nilai Mean Absolute Error (MAE), dan Mean Squared Error (MSE). Ketiga metode ini memberikan wawasan penting untuk pengambilan keputusan praktis. R^2 menunjukkan seberapa baik model menjelaskan variabilitas data, MAE mengukur kesalahan rata-rata dalam satuan yang sama dengan data, memudahkan interpretasi, dan MSE memberi penalti pada kesalahan besar, membantu mengidentifikasi ketidakkonsistenan model [11][10]. Kombinasi ketiga metrik ini memastikan akurasi model dan keandalannya dalam penerapan praktis.

6.1. Nilai koefisien determinasi atau R^2 menunjukkan keakuratan model. Nilai R^2 diperoleh dengan persamaan berikut.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Dengan:

SS_{res} : *Sum Squared Residual* (jumlah kuadrat galat)
SS_{tot} : *Sum Squared Total* (jumlah kuadrat total)
y_i : Nilai yang diamati
$\hat{y}$: Nilai hasil prediksi
$\bar{y}$: Rata-rata nilai yang diamati

Nilai R^2 merentang dari 0 hingga 1. Jika nilai R2 mendekati 1, maka model semakin akurat.

6.2. *Mean Absolute Error* (MAE) adalah nilai rata-rata dari absolut *error* dari nilai prediksi model. Nilai MAE diperoleh dengan persamaan berikut.

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

Dengan:

n : jumlah data
y_i : nilai hasil sebenarnya
$\hat{y}_i$: nilai dari hasil prediksi

6.3. *Mean Squared Error* (MSE) adalah nilai rata-rata dari hasil kuadrat error dari nilai prediksi model. Nilai MSE diperoleh dengan persamaan berikut.

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

Dengan:

n : jumlah data
y_i : nilai hasil sebenarnya
$\hat{y}_i$: nilai dari hasil prediksi

Semakin kecil nilai MAE dan MSE yang diperoleh, semakin baik model yang digunakan.

HASIL PENELITIAN

1. Pengumpulan Data

Data yang digunakan dalam penelitian kali ini diperoleh dengan metode dokumentasi. Pengumpulan data dengan metode ini dilakukan dengan mengunduh data berbentuk CSV yang diperoleh dari situs https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera. Berkas CSV tersebut berisi data produksi padi di Pulau Sumatera dari tahun 1993 hingga 2020 beserta faktor-faktor yang mempengaruhi tingkat produksi padi yang diperoleh dari situs resmi BPS. Deskripsi dataset berbentuk tabel yang berisi 224 baris dan 7 kolom. Data pada tahun 2021 dan 2022 diperoleh secara manual dari

situs BPS dan BMKG. Pada penelitian ini, data yang digunakan hanya dari tahun 2012 hingga 2022 dengan jumlah data sebanyak 88 baris dan 7 kolom.

Tabel 1 Cuplikan Data

Provinsi	Tahun	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
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
...	...	...	...	...	...	...
Lampung	2018	2488641.91	511940.93	1385.8	76.05	25.5
Lampung	2019	2164089.33	464103.42	1706.4	78.03	27.23
Lampung	2020	2604913.29	545149.05	2211.3	75.8	24.58
Lampung	2021	2485452.78	489573.23	2110.5	81.77	27.13
Lampung	2022	2688160	518256.1	1848.1	98.75	34.35

2. Preprocessing Data

2.1 Memeriksa Kolom dengan Nilai Kosong

Suatu dataset perlu diperiksa sebelum memasuki tahap pemodelan. Hal ini perlu dilakukan karena algoritma pada Python tidak dapat memproses data dengan nilai kosong.

```
data.isnull().sum()
```

	0
Provinsi	0
Tahun	0
Produksi	0
Luas Panen	0
Curah hujan	0
Kelembapan	0
Suhu rata-rata	0

Gambar 1 Jumlah Data Kosong

Setelah diperiksa, dapat dilihat bahwa tidak ada kolom yang tidak berisi suatu nilai, sehingga dapat lanjut ke proses berikutnya.

2.2 Memeriksa Data Duplikat

Setelah memeriksa data yang tidak berisi suatu nilai, data perlu diperiksa akan adanya data duplikat. Data duplikat perlu dihilangkan karena akan berpengaruh terhadap hasil pemodelan.

```
data.duplicated().sum()

0
```

Gambar 2 Jumlah Data Duplikat

Hasil pada Gambar 2 menunjukkan bahwa tidak ada data duplikat, sehingga pemodelan dapat dilanjutkan.

2.3 Statistika Deskriptif

Statistika deskriptif merupakan prosedur statistika yang memberikan gambaran ringkas dari data, sehingga data dapat mudah dipahami [12]. Dari proses statistika deskriptif, terdapat beberapa informasi yang dapat diperoleh mengenai data, beberapa di antaranya adalah banyaknya data, nilai rata-rata, standar deviasi, median, minimum, maksimum, dan kuartil.

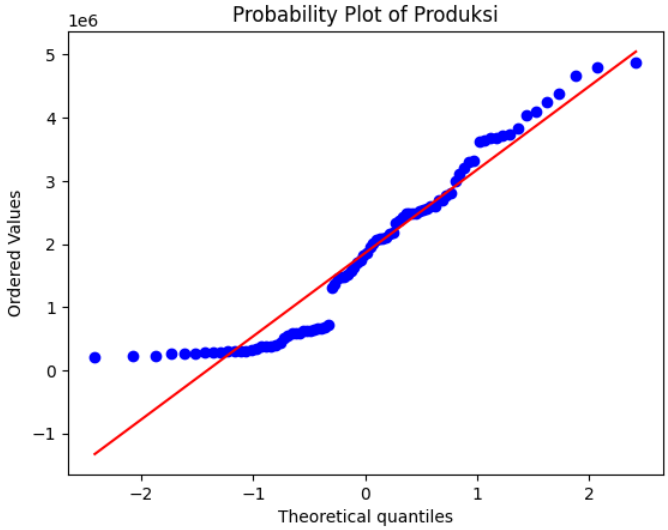
	Tahun	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
count	88.0000	8.800000e+01	88.000000	88.000000	88.000000	88.000000
mean	2017.0000	1.859705e+06	337651.889432	2576.829205	81.753750	26.768977
std	3.1804	1.348091e+06	233301.319641	1016.329507	3.700832	1.294610
min	2012.0000	2.135572e+05	51054.040000	327.330000	71.220000	22.190000
25%	2014.0000	5.693620e+05	107168.750000	1784.900000	79.575000	26.397500
50%	2017.0000	1.840815e+06	312361.025000	2330.300000	81.900000	26.835000
75%	2020.0000	2.690339e+06	504284.750000	3241.725000	83.855000	27.135000
max	2022.0000	4.881089e+06	872737.000000	5332.300000	98.750000	34.350000

Gambar 3 Statistika Deskriptif

Berdasarkan uraian pada Gambar 3, rata-rata produksi padi selama 10 tahun adalah 1.859.705 ton dengan produksi terendahnya sebanyak 2.135.572 ton dan produksi tertinggi sebanyak 4.881.089 ton. Pada Gambar 3, dapat dilihat juga bahwa nilai rata-rata dan median dari setiap variabel tidak berbeda jauh. Hal ini menunjukkan bahwa data cenderung berdistribusi normal.

2.4 Persebaran Data Variabel Target

Sebelum melakukan pemodelan regresi linear, persebaran data perlu diperiksa untuk memastikan data berdistribusi normal. Hal ini dilakukan untuk memastikan validitas persamaan regresi yang dihasilkan [13]. Selain dengan cara melihat kesamaan nilai rata-rata dan median, salah satu cara untuk memeriksa distribusi data adalah dengan QQ-plot. QQ-plot merupakan grafik yang membandingkan nilai-nilai pada kuantil suatu data dengan kuantil dari distribusi yang diharapkan [14]. Kuantil distribusi yang diharapkan adalah kuantil dengan distribusi normal. Dalam grafik, kuantil distribusi yang diharapkan direpresentasikan dengan garis diagonal. Jika nilai-nilai kuantil dari data yang diuji berada dekat pada garis diagonal, maka dapat disimpulkan bahwa data yang diuji berdistribusi normal [15].

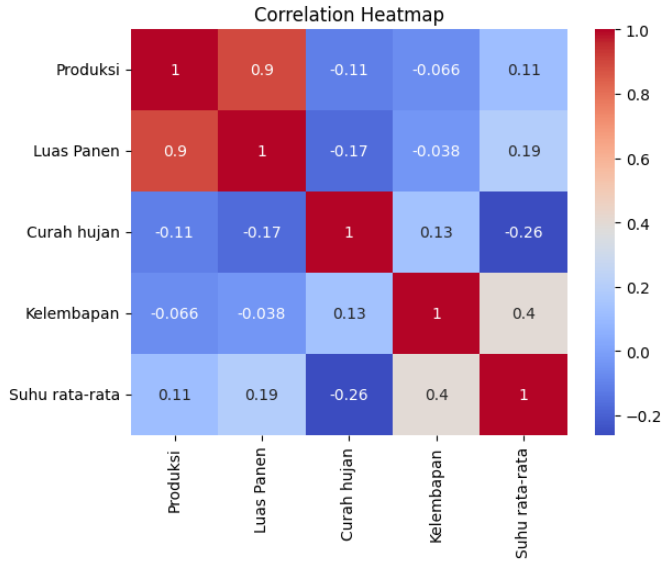


Gambar 4 QQ-Plot Variabel Target

Berdasarkan grafik pada Gambar 4, persebaran data cenderung dekat dengan garis diagonal, sehingga dapat disimpulkan bahwa data berdistribusi normal.

2.5 Matriks Korelasi

Matriks korelasi digunakan untuk menganalisis tingkat pengaruh variabel independen terhadap variabel dependen. Jika korelasi antar variabel bernilai kurang dari 0.5, maka korelasi variabel lemah. Namun, jika nilai korelasi bernilai 0.5 atau lebih, maka korelasi antar variabel kuat.

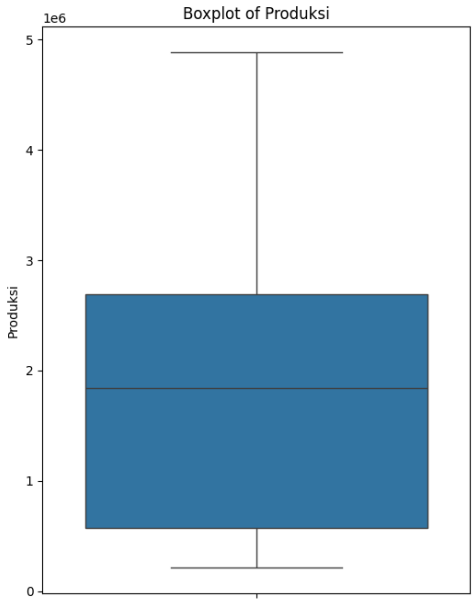


Gambar 5 Matriks Korelasi

Berdasarkan matriks korelasi yang telah dibuat, luas panen memiliki pengaruh yang signifikan terhadap jumlah produksi. Pada matriks tersebut, meskipun pengaruhnya tidak signifikan, dapat dilihat bahwa curah hujan dan kelembapan memiliki korelasi negatif terhadap produksi padi. Hal ini berarti jika nilai pada kedua variabel tersebut menurun, maka jumlah produksi padi akan bertambah.

2.6 Outlier

Outlier dapat didefinisikan sebagai pengamatan yang sangat berbeda atau jauh dari pengamatan lainnya. Kehadiran outlier dapat menyebabkan data tidak terdistribusi secara normal dan mempengaruhi kesimpulan atau keputusan yang diambil dalam penelitian [16]. Oleh karena itu, perlu dipastikan bahwa tidak ada outlier dalam data, sehingga hasil yang diperoleh lebih akurat dan representatif. Salah satu metode untuk memeriksa outlier adalah dengan menggunakan boxplot [16].



Gambar 6 Boxplot Variabel Target

Hasil boxplot pada Gambar 6 menunjukkan bahwa tidak ada outlier dalam data, yang berarti semua nilai dalam dataset berada dalam kisaran yang wajar dan tidak ada data yang menyimpang jauh dari nilai-nilai lainnya.

2.7 Pemilihan Fitur

Untuk melakukan pemodelan, variabel atau fitur didefinisikan ke dalam variabel x dan y agar algoritma dapat membedakan antara variabel independen dan independen.

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

Gambar 7 Variabel X dan Y

Variabel x berisi variabel-variabel independen yaitu luas panen, curah hujan, kelembapan, dan suhu rata-rata. Sedangkan variabel y berisi variabel dependen yaitu produksi padi.

2.8 *Feature Scaling*

Feature Scaling merupakan proses normalisasi data sehingga data berada dalam skala yang konsisten [9]. Dengan data berada dalam skala yang konsisten, dapat dipastikan bahwa variabel independen memiliki kontribusi yang setara terhadap variabel dependen atau target. Salah satu metode *feature scaling* adalah dengan metode *Standard Scaling*. *Standard Scaling* mengubah data sehingga data memiliki nilai rata-rata 0 dan standar deviasi 1 [9].

```
from sklearn.preprocessing import StandardScaler

scaler_x = StandardScaler()
scaler_y = StandardScaler()
x_scaled = scaler_x.fit_transform(x)
y_scaled = scaler_y.fit_transform(y.values.reshape(-1, 1))
print(x_scaled[:5])
print(y_scaled[:5])
```

```
[[ 0.21619471 -1.46340722 -0.5852988   0.10178628]
 [ 0.35146969 -0.94328845 -0.286365    0.17947225]
 [ 0.16590415 -0.30917103 -0.93858421  0.25715822]
 [ 0.53199582 -0.99138162 -0.4765956   0.25715822]
 [-0.19219948 -1.46538636  0.42564098  0.27269542]]
[[-0.2068859 ]
 [ 0.35163976]
 [-0.02957514]
 [ 0.07254143]
 [ 0.2395158 ]]
```

Gambar 8 *Standard Scaling* pada Dataset

3. *Train Test Split*

   Train Test Split membagi dataset menjadi data latih dan data uji. 80% jumlah data digunakan sebagai data latih, sedangkan 20% sisanya digunakan sebagai data uji.

```
x_train: (70, 4)
x_test:  (18, 4)
y_train: (70, 1)
y_test:  (18, 1)
```

Gambar 9 Jumlah Data Setelah *Train Test Split*

   Setelah pembagian, data yang digunakan untuk membangun model berjumlah 70 data, sedangkan data yang digunakan untuk menguji model berjumlah 18 data.

4. Pembuatan Model Regresi Linear

   Pemodelan regresi linear pada Python dapat dilakukan dengan mudah dengan mengimpor library sklearn.

```
from sklearn.linear_model import LinearRegression
linreg = LinearRegression()
linreg.fit(x_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

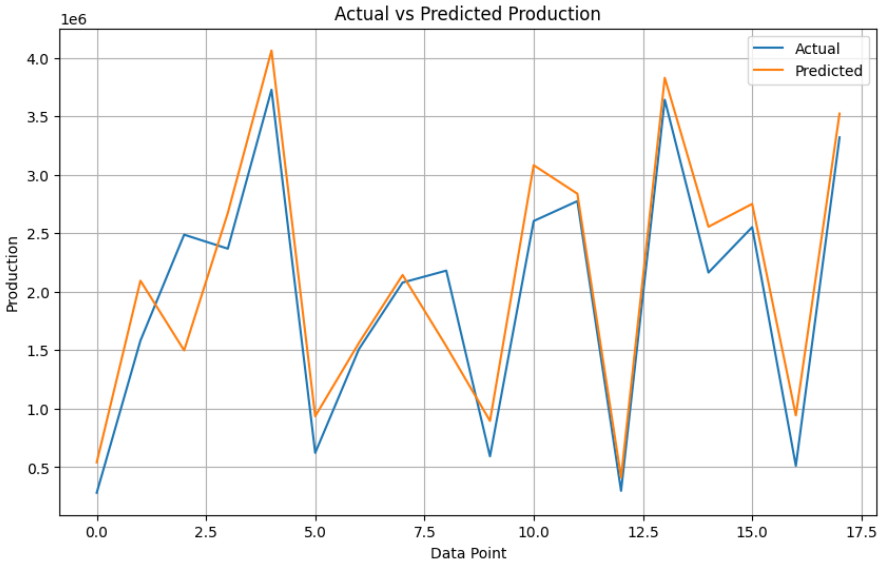
Gambar 10 Model Regresi Linear

   Setelah model dibangun dan dilatih dengan data latih, langkah berikutnya adalah melakukan pengujian terhadap model dengan data uji. Dikarenakan data telah distandarisasi nilainya dengan Standard Scaling, data uji dan hasil prediksi model terhadap data uji perlu dikembalikan ke dalam skala awal untuk dapat diketahui nilai sebenarnya.

```
y_pred_linreg = linreg.predict(x_test)
y_pred_linreg_inv = scaler_y.inverse_transform(y_pred_linreg)
y_test_inv = scaler_y.inverse_transform(y_test)
```

Gambar 11 Mengembalikan Hasil Prediksi Model Regresi Linear ke Skala Awal

Setelah dikembalikan ke skala aslinya, evaluasi dapat dilakukan dengan membandingkan nilai prediksi yang dihasilkan model dengan data uji menggunakan grafik.



Gambar 12 Perbandingan Hasil Prediksi Model Regresi Linear dengan Data Uji

Pada grafik perbandingan Gambar 12, dapat dilihat bahwa nilai yang diprediksi oleh model tidak jauh berbeda dengan nilai pada data uji. Hal ini menunjukkan bahwa model memiliki akurasi yang cukup tinggi.

	Actual	Predicted (LinReg)
0	281610.10	541902.55
1	1582393.00	2094751.73
2	2487929.00	1498759.78
3	2368390.00	2674255.82
4	3727249.00	4062353.72
5	622832.00	936105.21
6	1509456.00	1562611.94
7	2078901.59	2143484.64
8	2180754.00	1534451.23
9	593194.00	896229.37
10	2604913.29	3081954.43
11	2775069.00	2838146.58
12	298149.25	412572.36
13	3641895.00	3829267.54
14	2164089.33	2555832.96
15	2552443.19	2751020.21
16	512152.00	942982.93
17	3320064.00	3523263.76

Gambar 13 Tabel Perbandingan Hasil Prediksi dengan Data Uji

Pada Gambar 13, dapat dilihat bahwa hasil prediksi model regresi linear tidak berbeda signifikan dengan data sebenarnya. Namun, terdapat beberapa titik yang hasilnya kurang akurat, seperti yang dapat dilihat pada baris 2 dan 8 dari Gambar 13.

5. Pembuatan Model KNN

Sebelum melakukan pemodelan KNN dalam Python, perlu analisis untuk menentukan nilai k yang paling optimal. Analisis tersebut dapat dilakukan dengan membandingkan nilai error pada prediksi data latih dan data uji menggunakan model KNN dengan nilai k yang berbeda.

Gambar 14 Perbandingan Error Hasil Prediksi Data Latih dan Data Uji

Berdasarkan Gambar 14, nilai k yang paling optimal untuk digunakan adalah 6. Hal ini dikarenakan di saat k bernilai 6, model berhasil memprediksi data uji dengan nilai error yang terendah dibanding nilai lainnya. Selain itu, pada nilai k tersebut, nilai error pada prediksi data latih tidak berbeda jauh dengan nilai error pada prediksi data uji.

Setelah nilai k yang paling optimal ditentukan, model dibangun dengan mengimpor library sklearn dan membangun data dengan data latih.

```
from sklearn.neighbors import KNeighborsRegressor
knn = KNeighborsRegressor(n_neighbors=6)
knn.fit(x_train, y_train)
```

```
KNeighborsRegressor
KNeighborsRegressor(n_neighbors=6)
```

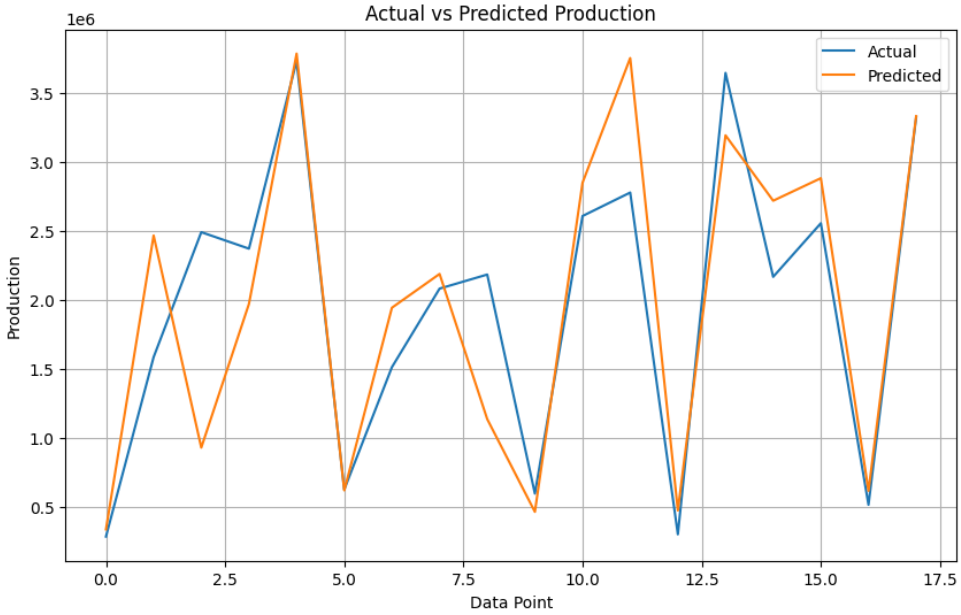
Gambar 15 Model KNN

Setelah model dibangun, model diuji dengan data uji dan nilainya dikembalikan ke skala aslinya untuk digunakan dalam evaluasi.

```
y_pred_knn = knn.predict(x_test)
y_pred_knn_inv = scaler_y.inverse_transform(y_pred_knn)
y_test_inv = scaler_y.inverse_transform(y_test)
```

Gambar 16 Mengembalikan Hasil Prediksi Model KNN ke Skala Awal

Data uji dan hasil prediksi model yang telah dikembalikan ke aslinya kemudian dibandingkan dalam bentuk grafik.



Gambar 17 Perbandingan Hasil Prediksi Model KNN dengan Data Uji

Grafik pada Gambar 17 menunjukkan model KNN berhasil memprediksi dengan cukup akurat. Namun, terdapat beberapa perbedaan yang cukup signifikan dalam beberapa titik data.

	Actual	Predicted (KNN)
0	281610.10	333647.66
1	1582393.00	2463872.25
2	2487929.00	926457.75
3	2368390.00	1967896.79
4	3727249.00	3781890.64
5	622832.00	616342.68
6	1509456.00	1940206.64
7	2078901.59	2185080.93
8	2180754.00	1132968.67
9	593194.00	460930.90
10	2604913.29	2846324.70
11	2775069.00	3749804.67
12	298149.25	470978.39
13	3641895.00	3189758.31
14	2164089.33	2715724.81
15	2552443.19	2879120.25
16	512152.00	613181.91
17	3320064.00	3328640.23

Gambar 18 Tabel Perbandingan Hasil Prediksi dengan Data Uji

Titik data berbeda signifikan dapat dilihat pada Gambar 18. Beberapa di antaranya ada pada baris 1, 2, dan 8.

6. Evaluasi Model

Berdasarkan uraian sebelumnya, model regresi linear maupun KNN mampu memprediksi nilai dengan cukup akurat. Selain membandingkan kemiripan hasil prediksi dengan data uji menggunakan grafik, akurasi model dapat diukur dengan membandingkan nilai R^2, MAE dan MSE. Hasil perhitungan evaluasi kedua model tersebut adalah sebagai berikut.

	Model	R2 Score	MAE	MSE
0	Linear Regression	0.868181	324967.053821	1.571551e+11
1	KNN	0.703748	416812.345648	3.531935e+11

Gambar 19 Perbandingan Nilai R^2, MAE dan MSE Model

Berdasarkan Gambar 19, model regresi linear menunjukkan performa yang lebih unggul dibandingkan KNN, dengan nilai R^2 yang lebih tinggi serta MAE dan MSE yang lebih rendah. Ini menunjukkan bahwa regresi linear mampu menangkap pola hubungan antara variabel secara lebih efektif dibandingkan KNN, yang cenderung sensitif terhadap distribusi data dan parameter k yang digunakan. Dalam konteks memprediksi produksi padi di Pulau Sumatera, akurasi model sangat penting untuk memberikan estimasi yang andal bagi pengambilan keputusan, seperti perencanaan distribusi hasil panen, pengelolaan stok, dan strategi pengembangan pertanian. Dengan performa yang lebih konsisten, regresi linear memberikan dasar yang lebih kuat untuk memahami tren dan pola produksi, sehingga mendukung upaya optimalisasi sektor pertanian di wilayah ini.

7. Implementasi Model

Sebagai pembuktian lebih lanjut, model akan diimplementasikan untuk memprediksi produksi padi dengan data yang belum digunakan sebelumnya dalam proses pembangunan maupun pengujian model. Data yang digunakan adalah data produksi padi pada 2023 di Provinsi Sumatera Selatan yang diperoleh dari situs BPS dan BMKG.

```
datanew = {'Produksi': 2832770, 'Luas Panen': 504140, 'Curah hujan': 27.88, 'Kelembapan': 81.32, 'Suhu rata-rata': 10.42}
datanew = pd.DataFrame([datanew])
x_new = datanew.drop(columns='Produksi', axis=1)
y_new = datanew['Produksi']
```

Gambar 20 Data Baru

Setelah mendefinisikan nilai x dan y dari data yang baru, data distandarisasi dengan Standard Scaling lalu nilai x diprediksi menggunakan kedua model.

```
x_new_scaled = scaler_x.transform(x_new)
y_new_scaled = scaler_y.transform(y_new.values.reshape(-1, 1))

y_pred_linreg_scaled = linreg.predict(x_new_scaled)
y_linreg_pred = scaler_y.inverse_transform(y_pred_linreg_scaled.reshape(-1, 1))

y_pred_knn_scaled = knn.predict(x_new_scaled)
y_knn_pred = scaler_y.inverse_transform(y_pred_knn_scaled.reshape(-1, 1))

print(f'Actual: {datanew["Produksi"][0]}')
print(f'Predicted Linear Regression: {y_linreg_pred[0][0]}')
print(f'Predicted KNN: {y_knn_pred[0][0]}')

Actual: 2832770
Predicted Linear Regression: 2822234.1693238975
Predicted KNN: 2003718.2583333333
```

Gambar 21 Hasil Prediksi Data Baru

Setelah nilai dikembalikan ke skala semula, hasilnya kemudian dibandingkan dengan data asli. Perbandingan tersebut menunjukkan bahwa prediksi menggunakan regresi linear hampir menyerupai nilai aktual.

PEMBAHASAN

Penelitian ini bertujuan untuk membandingkan kinerja dua metode prediksi, yaitu regresi linear berganda dan K-Nearest Neighbor (KNN), dalam meramalkan produksi padi di Pulau Sumatera. Regresi linear berganda merupakan metode yang digunakan untuk menganalisis hubungan antara beberapa variabel independen dengan satu variabel dependen [12], sementara KNN memprediksi nilai berdasarkan kedekatannya dengan data yang ada [7]. Sebelumnya, model prediksi menggunakan regresi linear telah dikembangkan, namun penelitian ini bertujuan untuk mengevaluasi apakah KNN dapat menghasilkan prediksi yang lebih akurat untuk produksi padi di wilayah tersebut.

Penelitian ini mempertimbangkan beberapa faktor yang memengaruhi produksi padi, namun hanya fokus pada beberapa faktor utama, seperti luas panen, curah hujan, kelembapan, dan suhu rata-rata. Wilayah penelitian terbatas pada Pulau Sumatera, yang mencakup Provinsi Nanggroe Aceh Darussalam (NAD), Sumatera Utara, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Lampung.

Hasil penelitian menunjukkan bahwa regresi linear lebih akurat dalam memprediksi produksi padi dibandingkan dengan KNN. Model regresi linear menunjukkan prediksi yang lebih mendekati nilai aktual, sementara KNN cenderung memiliki prediksi yang sedikit lebih variatif. Hal ini dapat dibuktikan dengan nilai R^2 model regresi linear yang mencapai 0.868181, lebih tinggi dibandingkan dengan nilai R^2 pada model KNN yang sebesar 0.703748, dengan selisih sebesar 18,94%. Selain itu, model regresi linear juga menunjukkan performa yang lebih baik dalam hal kesalahan prediksi, dengan nilai MAE sebesar 324967.053821, yang lebih rendah 22,03% dibandingkan model KNN. Begitu pula, nilai MSE pada model regresi linear yang sebesar 1.571551e+11 lebih rendah 55,49% dibandingkan dengan model KNN.

Model yang dikembangkan dalam penelitian ini memiliki keunggulan dalam memprediksi produksi padi di Pulau Sumatera. Namun, jika model ini ingin diterapkan di wilayah lain, perlu diperhatikan bahwa perbedaan karakteristik wilayah, seperti kondisi tanah, iklim, dan pola tanam, dapat memengaruhi akurasi prediksi. Oleh karena itu, langkah-langkah seperti pengumpulan data lokal, pelatihan ulang model, dan evaluasi kinerjanya menjadi sangat penting untuk memastikan model tetap relevan dan akurat sesuai dengan kondisi spesifik wilayah tersebut. Selain itu, penting untuk terus memperbarui data yang digunakan dalam model. Pola produksi padi dapat berubah dari waktu ke waktu akibat faktor seperti perubahan iklim, adopsi teknologi baru, atau kebijakan agrikultur yang berbeda. Dengan memasukkan data terbaru, model tidak hanya mampu mengenali tren baru tetapi juga memberikan prediksi yang lebih relevan dan andal, sehingga mendukung pengambilan keputusan yang lebih tepat di masa depan.

Untuk peneliti selanjutnya diharapkan dapat meneliti menggunakan metode dan variabel yang berbeda untuk menambah referensi yang lebih baik bagi siapa saja yang membutuhkan mengenai penelitian terkait.

KESIMPULAN

Penelitian ini membandingkan kinerja model regresi linear dan K-Nearest Neighbor (KNN) dalam prediksi produksi tanaman padi di Pulau Sumatera. Hasil evaluasi menunjukkan bahwa regresi linear lebih unggul dalam hal akurasi dalam memprediksi dengan kesalahan prediksi yang lebih rendah dibandingkan KNN.

Penggunaan regresi linear dalam penelitian ini mencapai nilai R^2 sebesar 0,868181, yang lebih tinggi 18,94% dibandingkan KNN yang memiliki nilai R^2 sebesar 0,703748. Selain itu, nilai MAE pada regresi linear adalah 324967,053821, lebih rendah 22,03% dibandingkan nilai MAE KNN sebesar 416812,345648. Begitu pula, nilai MSE pada regresi linear sebesar 1,571551e+11 lebih rendah 55,49% dibandingkan nilai MSE KNN sebesar 3,531935e+11. Hasil dari penelitian ini dapat dimanfaatkan sebagai landasan dalam pengambilan keputusan strategis di sektor pertanian, khususnya untuk merencanakan dan mengelola produksi padi di Pulau Sumatera. Dengan demikian, model prediksi yang dihasilkan dapat berfungsi sebagai alat yang andal untuk mendukung perencanaan pertanian yang lebih efisien dan terstruktur.

REFERENCES

[1] Y. Nababan and I. Nugraha, "Penerapan Data Mining Produksi Padi di Pulau Sumatera Menggunakan Analisis Regresi Linear", *JUTIN*, vol. 7, no. 1, pp. 262–272, 2024, doi: https://doi.org/10.31004/jutin.v7i1.23545.

[2] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, "Penerapan Algoritme Linear Regression untuk

Prediksi Hasil Panen Tanaman Padi", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 8, no. 4, pp. 364-370, 2019, https://journal.ugm.ac.id/v3/JNTETI/article/view/2563 (accessed Maret 3, 2024).

[3] M. D. Irrawati and M. Mukaramah, "Implementasi Metode Regresi Linear Berganda untuk Mengatasi Pelanggaran Asumsi Klasik ", *Studi Akuntansi, Keuangan dan Manajemen*, vol. 3, no. 2, pp. 83–94, 2024, doi: https://doi.org/10.35912/sakman.v3i2.2743.

[4] E. Mardiani et al., "Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma", *Innovative*, vol. 3, no. 2, pp. 13880–13892, 2023, https://j-innovative.org/index.php/Innovative/article/view/1949 (accessed Mar 3, 2024).

[5] R. D. Ayu, "11 Daerah Penghasil Padi Terbesar di Indonesia", *koran.tempo.co*, 2023, https://koran.tempo.co/read/berita-utama/485632/11-daerah-penghasil-padi-terbesar-di-Indonesia (accessed Mar 4, 2024).

[6] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara", *ijodas*, vol. 1, no. 2, pp. 39-43, Jul. 2020. doi: https://doi.org/10.33096/ijodas.v1i2.13.

[7] D. S. Seruni, M. T. Furqon, dan R. C. Wihandika, "Sistem Prediksi Pertumbuhan Jumlah Penduduk Kota Malang menggunakan Metode K-Nearest Neighbor Regression", *J-PTIIK*, vol. 4, no. 4, pp. 1075–1082, 2020, https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/7135 (accessed Mar. 04, 2024).

[8] E. Triyanto, H. Sismoro, and A. Laksito, "IMPLEMENTASI ALGORITMA REGRESI LINEAR BERGANDA UNTUK MEMPREDIKSI PRODUKSI PADI DI KABUPATEN BANTUL", *rabit*, vol. 4, no. 2, pp. 73-86, Jul. 2019. doi: https://doi.org/10.36341/rabit.v4i2.666.

[9] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning", *JAETS*, vol. 5, no. 1, pp. 401–413, 2023, doi: https://doi.org/10.37385/jaets.v5i1.3080.

[10] I. Amansyah, J. Indra, E. Nurlaelasari, and A. R. Juwita, "Prediksi Penjualan Kendaraan Menggunakan Regresi Linear: Studi Kasus pada Industri Otomotif di Indonesia", *Innovative*, vol. 4, no. 4, pp. 1199–1216, 2024, doi: https://doi.org/10.31004/innovative.v4i4.12735.

[11] A. R. R. Aditia, M. Wadud, and M. K. DP, "Pengaruh Kualitas Produk terhadap Kepuasan Konsumen Sepeda Motor NMAX pada PT Yamaha A. Rivai Palembang", JNMPSDM, vol. 1, no. 1, pp. 23-37, Sep. 2020., doi: https://doi.org/10.47747/jnmpsdm.v1i01.4.

[12] M. A. Musababa, "Implementasi Algoritma Linear Regression untuk Prediksi Produksi Tanaman Padi di Kabupaten Grobogan", *Data Sci. Indones.*, vol. 3, no. 2, pp. 68-78, 2023, doi: 0.47709/dsi.v3i2.3118.

[13] M. Tarigan and D. F. Silaban, "Statistika Deskriptif", *JINTAN: Jurnal Ilmu Keperawatan*, vol. 4, no. 2, pp. 187-195, 2024, doi: https://doi.org/10.51771/jintan.v4i2.859.

[14] A. S. P. Pramono and A. Ahdika, "Analisis Regresi Berganda pada Faktor-Faktor yang Mempengaruhi Kinerja Fisik Preservasi Jalan dan Jembatan Di Provinsi Sumatera Selatan", *Emerg. Stat. and Data Sci. J.*, vol. 1, no. 1, pp. 47-56, 2022, doi: https://doi.org/10.20885/esds.vol1.iss.1.art6

[15] P. Guzik and B. Więckowska, "Data distribution analysis – a preliminary approach to quantitative data in biomedical research", *JMS*, vol. 92, no. 2, p. e869, Jun. 2023, doi: 10.20883/medical.e869.

[16] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya ", *JESI*, vol. 2, no. 3, pp. 307-316, Jan. 2023, doi: 10.11594/jesi.02.03.07.