Laporan Akhir Tugas MK. Data Mining (KOM1338), Semester Genap 2023/2024

Implementasi Random Forest Regression untuk Prediksi Hasil Panen Tanaman Pangan di Sumatera

Halida Fiadnin (G6401211142), Ismy Fana Fillah (G64012), Tita Madriyanti (G6401211120), Viragita Athaya Haura (G6401211116) [1*]
Kelompok: 4, Kelas Paralel: 1

Abstrak

Abstrak ditulis dalam 1 paragraf dan panjangnya tidak lebih dari 200 kata. Abstrak dimulai dengan uraian latar belakang tugas akhir dalam 2-3 kalimat, metode, dan hasil temuan utama yang secara langsung menjawab masalah yang dikaji. Hindari penggunaan singkatan.

Kata Kunci**:** Kata Kunci terdiri atas maksimum 5 kata yang diurutkan mengikuti abjad.

PENDAHULUAN

Latar Belakang

Sebagai negara agraris, Indonesia menghasilkan berbagai bahan pokok pertanian seperti beras, jagung, ubi jalar, kacang tanah, serta berbagai bahan pokok lainnya. Hal ini menandakan bahwa pertanian menjadi salah satu sektor utama dalam kebutuhan pangan Indonesia. Pulau Sumatera, sebagai salah satu wilayah paling produktif, berperan penting dalam memenuhi kebutuhan pangan pokok tersebut (Satria *et al.* 2023). Iklim di Pulau Sumatera sangat mendukung aktivitas pertanian karena curah hujan yang merata sepanjang tahun (Sumaryanto 2012). Hal ini diperkuat oleh kondisi geografis deretan Bukit Barisan yang memudahkan terbentuknya hujan orografis (Isbah dan Iyan 2016). Akan tetapi, Perubahan iklim cuaca yang tidak stabil berpengaruh besar pada hasil pertanian dapat menyebabkan kerugian besar dalam sektor pertanian karena pola cuaca yang tidak terduga (Fitri dan Nugraha 2024). Dampak perubahan iklim terhadap sektor pertanian bergantung pada tingkat dan laju perubahan iklim. Perubahan iklim menyebabkan peningkatan frekuensi dan intensitas cuaca ekstrem, perubahan pola hujan, serta kenaikan suhu dan permukaan air laut. Cuaca ekstrem ini dapat mengakibatkan kegagalan dalam panen dan penanaman, yang berujung pada penurunan produktivitas dan produksi serta kerusakan pada sumber daya lahan pertanian. (Nuraisah dan Kusumo 2019). Selain itu, jenis tanah dan topografi wilayah memiliki peran penting dalam menentukan hasil panen (Herlina dan Prasetyorini 2020), karena berbagai karakteristik tanah, seperti kesuburan dan kemampuan menyimpan air, serta kontur dan kemiringan lahan, dapat mempengaruhi produktivitas tanaman.

Dalam mendukung hasil lahan yang baik penunjang pertanian presisi, perlu dilakukan implementasi teknik data mining sebagai model prediksi untuk meningkatkan produktivitas dan ketahanan pangan. Model ini juga sangat penting dalam membantu petani dan pemangku kepentingan terkait untuk membuat keputusan guna mengoptimalkan hasil panen. Prediksi hasil pertanian komoditas pangan sangat dipengaruhi oleh perubahan iklim yang berkaitan dengan curah hujan, kelembapan, suhu, dan lain sebagainya. Oleh karena itu, variasi iklim cuaca tahunan menjadi variabel independen yang dapat

[1]Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: username1@yahoo.co.id, username1@yahoo.co.id, viragitaathayahaura@apps.ipb.ac.id

mempengaruhi hasil panen komoditas tanaman pangan di Sumatera (Satria et al. 2023). Penelitian ini melakukan prediksi hasil pertanian komoditas tanaman pangan, termasuk padi, jagung, kacang tanah, kedelai, ubi kayu, dan ubi jalar di pulau Sumatera, dengan menggunakan pendekatan teknik data mining model regresi algoritma Random Forest Regression (RFR). Random Forest Regression merupakan teknik dalam machine learning yang menggunakan konsep ensemble learning, khususnya sebagai model regresi. Dalam ensemble learning, hasil dari berbagai model digabungkan untuk meningkatkan kinerja dan akurasi prediksi dibandingkan dengan hanya menggunakan satu model. Teknik ini juga menerapkan pengujian dengan konsep supervised learning dalam membangun kelas classifier, dengan menggabungkan prediksi dari beberapa Decision Tree (Fitri 2023). Analisis model regresi RFR dilakukan dengan mengukur nilai Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan R2 Score. RMSE dan MAE digunakan sebagai indikator untuk mengukur besarnya kesalahan prediksi dan seberapa baik model mencerminkan variasi data yang sebenarnya (Fitri dan Nugraha 2024).

Data yang digunakan pada penelitian ini merupakan data yang bersumber dari Badan Pusat Statistik BPS) memaparkan kondisi iklim cuaca lahan pertanian berbagai di pulau Sumatera dengan berbagai variabel yang berkaitan terhadap produksi dan produktivitas panen. Penelitian ini bertujuan sebagai implementasi teknik data mining model regresi algoritma Random Forest Regression (RFR) dalam memprediksi hasil pertanian komoditas tanaman pangan, termasuk padi, jagung, kacang tanah, kedelai, ubi kayu, dan ubi jalar di pulau Sumatera dalam menunjang pertanian presisi. Dengan dilakukannya penelitian ini, diharapkan dapat memberikan kontribusi signifikan terhadap peningkatan produktivitas pertanian dan ketahanan pangan di Sumatera. Melalui memanfaatkan model regresi Random Forest Regression (RFR), petani dan pemangku kepentingan diharapkan dapat membuat keputusan yang lebih tepat waktu dan akurat dalam pengelolaan lahan pertanian mereka. Selain itu, penelitian ini juga digunakan untuk menganalisis dan memahami lebih dalam tentang pengaruh variabel iklim terhadap hasil panen komoditas tanaman pangan di Sumatera. Dengan menggunakan data iklim yang komprehensif dari BPS berbagai variabel lain yang relevan, model RFR diharapkan mampu mengidentifikasi pola dan tren yang signifikan, sehingga dapat memberikan prediksi yang lebih akurat.

Tujuan

Berdasarkan latar belakang tersebut, tujuan penelitian dilakukan adalah sebagai berikut.

1. Memprediksi hasil pertanian komoditas seperti padi, jagung, kacang tanah, kedelai, ubi kayu, dan ubi jalar menggunakan teknik data mining model regresi Random Forest Regression (RFR).
2. Menganalisis dan memahami pengaruh variabel iklim terhadap hasil panen komoditas tanaman pangan di Sumatera.

Ruang Lingkup

Ruang lingkup penelitian yang dilakukan meliputi data dan teknik yang digunakan sebagai berikut.

1. Penelitian difokuskan pada berbagai provinsi di Pulau Sumatera, yaitu Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, dan Kepulauan Riau.
2. Menggunakan data iklim kondisi cuaca yang diperoleh dari Badan Pusat Statistik (BPS) dengan memiliki berbagai atribut seperti jenis tanaman, provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, suhu rata-rata, produktivitas, luas provinsi, dan rasio penggunaan lahan di Pulau Sumatera.

3. Melibatkan produksi komoditas pertanian utama seperti padi, jagung, kacang tanah, kedelai, ubi kayu, dan ubi jalar dari tahun 1993 hingga 2018.
4. Menggunakan data iklim kondisi cuaca yang diperoleh dari Badan Pusat Statistik (BPS), seperti curah hujan, kelembapan, suhu rata-rata.
5. Menerapkan implementasi data mining model regresi Random Forest Regression (RFR) untuk prediksi.
6. Model dievaluasi dengan mengukur nilai MAE, MSE, RMSE, dan R2 Score

Manfaat

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut.

1. Memberikan kontribusi signifikan terhadap peningkatan produktivitas pertanian dan ketahanan pangan di Sumatera.
2. Membantu petani dan pemangku kepentingan dalam membuat keputusan yang lebih tepat waktu dan akurat dalam pengelolaan lahan pertanian.
3. Menyediakan wawasan yang lebih dalam mengenai pengaruh variabel iklim terhadap hasil panen, sehingga dapat mengantisipasi dampak perubahan iklim.
4. Mengimplementasikan data mining dalam sektor pertanian untuk menunjang pertanian presisi.

TINJAUAN PUSTAKA

A. Regresi

Regresi adalah suatu metode statistik untuk membuat prediksi dari nilai yang ada pada satu variabel independen (prediktor) dengan mempertimbangkan nilai yang berada pada variabel lain (dependen atau kriteria). Tujuannya bukanlah untuk membuat prediksi yang sempurna, melainkan untuk membuat prediksi nilai variabel dependen dengan error yang sekecil-kecilnya (Trianggana 2020). Regresi berguna untuk mengetahui seberapa besar variabel prediktor mampu menjelaskan variasi dalam variabel dependen kriteria (Widhiarso 2010). Persamaan yang menyatakan bentuk hubungan antara variabel prediktor dan variabel kriteria dinyatakan dengan persamaan regresi atau model regresi. Model regresi dapat berupa hubungan linier, hubungan kuadratik, eksponen dan lainnya (Gultom *et al.* 2017)

B. Random Forest Regression

Random forest merupakan metode penggabungan atau bagging dengan cara membangkitkan sejumlah tree dari data sampel. Random forest akan memberikan pengukuran setiap variabel terpenting untuk digunakan sebagai prediktor agak mendapatkan akurasi yang tinggi (Satria *et al*. 2023). Random forest adalah kumpulan pohon prediktor, serta pengaturannya terhadap regresi (bukan klasifikasi) yang menjadikan hasilnya berupa numerik (Segal 2004). Random forest regression merupakan metode machine learning dengan konsep supervised, yaitu melakukan konsep *decision-tree* berulang kali hingga membentuk seolah *forest* atau hutan. Algoritma ini memodelkan prediksi dengan cara *multiple decision tree* (Saadah dan Salsabila 2021).

C. Mean Squared Error

Mean Squared Error (MSE) merupakan matriks yang digunakan untuk mengukur seberapa bagus model regresi dapat memprediksi nilai yang sebenarnya (Musababa 2023). MSE digunakan untuk mengevaluasi metode peramalan. Nilai pada MSE diperoleh dari selisih antara nilai yang sebenarnya dengan nilai yang

diprediksi oleh model lalu dikuadratkan. Pendekatan MSE ini terkadang menghasilkan jumlah yang besar dikarenakan selisih tersebut dikuadratkan. (Sautomo dan Pardede 2021). Nilai MSE yang semakin rendah menandakan bahwa kinerja model semakin baik (Zainal 2024).

D. Mean Absolute Error

Mean Absolute Error (MAE) merupakan matriks evaluasi alternatif yang digunakan untuk mengukur kesalahan prediksi dengan cara yang lebih sederhana dari MSE (Musababa 2023). Nilai pada MAE diperoleh dari rataan selisih absolut nilai yang sebenarnya dengan nilai yang diprediksi oleh model (Sautomo dan Pardede 2021). Nilai MAE yang semakin rendah juga menandakan bahwa kinerja model semakin baik (Zainal 2024).

E. Root Mean Squared Error

Root Mean Squared Error (RMSE) merupakan sebuah metode yang digunakan untuk mengukur tingkat kesalahan (error) suatu pelatihan (Marutho 2019). RMSE digunakan sebagai pembeda antara suatu nilai yang diprediksi dengan nilai yang sebenarnya (Sautomo dan Pardede 2021). RMSE diperoleh dengan menghitung akar kuadrat dari MSE, sehingga RMSE memberikan gambaran intuitif tentang kesalahan prediksi (Zainal 2024). Nilai RMSE yang semakin kecil menandakan bahwa hasil pengukurannya semakin akurat. Batas nilai RMSE adalah 0, yang artinya, semakin mendekati 0 maka nilai RMSE semakin akurat (Arinal dan Azhari 2023).

F. R-Squared Score

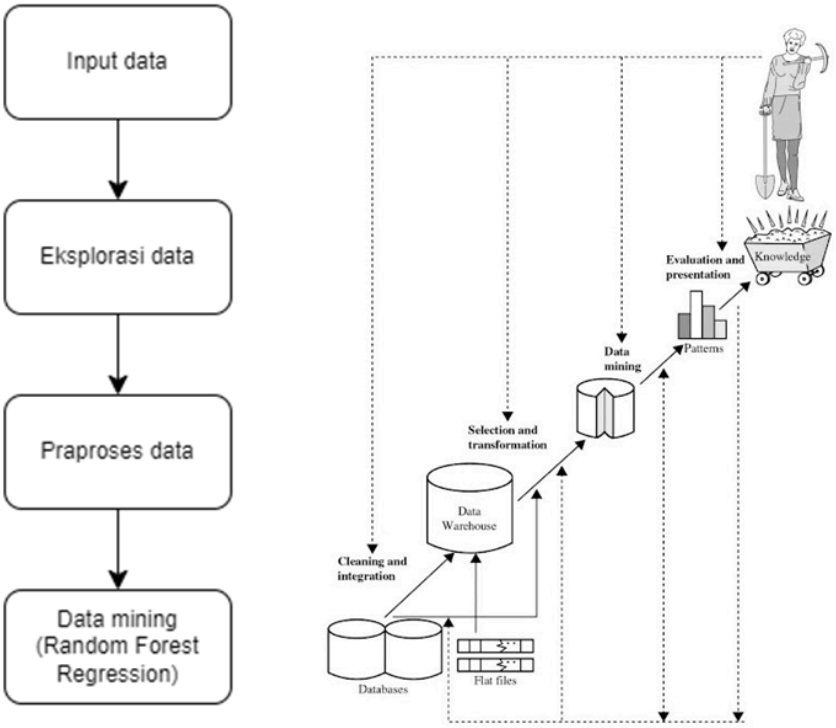
R-Squared Score (R2 Score) merupakan matriks yang digunakan untuk mengukur sejauh mana model regresi cocok dengan data. R2 akan mengukur proporsi varians yang ada dalam variabel dependen, yang dapat dijelaskan oleh variabel independen dalam model (Musababa 2023). Rentang nilai untuk R2 score adalah 0 hingga 1. Semakin mendekati angka 1 dan menjauhi angka 0, maka hasilnya semakin akurat (Hutahaean *et al*. 2024).

METODE

Data

Data yang digunakan dalam penelitian ini berasal dari Badan Pusat Statistik (BPS) yang menyediakan informasi mengenai kondisi iklim, cuaca, dan lahan pertanian di berbagai provinsi di Pulau Sumatera. Data ini mencakup berbagai atribut yang relevan untuk analisis prediksi hasil pertanian, yaitu jenis tanaman, provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, suhu rata-rata, produktivitas, luas provinsi, dan rasio penggunaan lahan. Data yang dikumpulkan mencakup periode dari tahun 1993 hingga 2018. Data disimpan dalam format CSV (*Comma-Separated Values*) yang memudahkan proses pembacaan dan pengolahan menggunakan perangkat lunak analisis data seperti Python.

Tahapan Kegiatan



Gambar 1 Tahapan penelitian mengacu knowledge discovery in database (KDD)

Pada tahapan awal penelitian ini dimulai dengan melakukan input data, data yang digunakan dalam penelitian ini diperoleh dari Badan Pusat Statistik (BPS), yang menyediakan informasi terkait kondisi iklim, cuaca, dan lahan pertanian di berbagai provinsi di Pulau Sumatera untuk periode 1993 hingga 2018. Data ini mencakup atribut-atribut penting seperti jenis tanaman, provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, suhu rata-rata, produktivitas, luas provinsi, dan rasio penggunaan lahan. Proses analisis dimulai dengan eksplorasi data untuk memahami struktur dan karakteristiknya melalui visualisasi dan statistik deskriptif. Sebelum dilakukan eksplorasi, *cleaning data* terlebih dahulu terhadap dataset untuk mengatasi *missing values*, outlier, dan duplikasi pada data sehingga data siap untuk dieksplorasi.

Setelah dilakukan eksplorasi terhadap data yang sudah di-*cleaning*, kemudian dilakukan reduksi data dengan seleksi fitur relevan dan reduksi dimensi. Pada tahapan praproses data ini, normalisasi data tidak dilakukan karena Random Forest tidak sensitif terhadap skala fitur. Algoritma ini menggunakan pohon keputusan yang membagi data berdasarkan nilai fitur relatif, bukan skala absolutnya. Metode pembagian dalam pohon keputusan Random Forest ditentukan oleh informasi gain atau indeks Gini, yang bekerja dengan membandingkan nilai fitur secara langsung. Oleh karena itu, variasi skala tidak mempengaruhi performa model secara signifikan. Setelah data melalui tahap praproses dan siap untuk diproses, dilakukan pembuatan model dari data yang dimiliki dengan mengimplementasikan teknik data mining model regresi Random Forest Regression. Analisis dilakukan menggunakan Random Forest Regression, sebuah teknik *ensemble learning* yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi hasil pertanian. Tahapan-tahapan ini memastikan data siap digunakan dan model dapat memberikan prediksi yang lebih akurat.

Lingkungan Pengembangan

Spesifikasi perangkat keras dan perangkat lunak yang digunakan dalam mendukung penelitian ini adalah sebagai berikut.

1. Perangkat keras
   a. Processor Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40 GHz
   b. RAM 8 GB
2. Perangkat lunak
   a. Sistem operasi : Windows 10
   b. Bahasa pemrograman : python
   c. Library : Pandas untuk manipulasi data, NumPy untuk komputasi numerik, Scikit-learn untuk implementasi algoritma machine learning, Matplotlib dan Seaborn untuk visualisasi data, dan berbagai library lainnya.
   d. Integrated Development Environment (IDE): Visual Studio Code atau Google Colab untuk pengembangan dan eksekusi kode.

HASIL DAN PEMBAHASAN

a. Import *Library* dan Dataset

```
!pip install seaborn
!pip install plotly

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Baca dataset
da = pd.read_csv('data_lengkap.csv')
```

Pada tahap ini, beberapa library digunakan untuk mendukung analisis dan visualisasi data. Library `pandas` (`import pandas as pd`) digunakan untuk membaca dan memanipulasi data dalam bentuk tabel atau DataFrame. Library `numpy` (`import numpy as np`) digunakan untuk komputasi numerik dan operasi array. Library `seaborn` (`import seaborn as sns`) digunakan sebagai antarmuka tingkat tinggi untuk membuat grafik statistik yang menarik dan informatif. Library `matplotlib.pyplot` (`import matplotlib.pyplot as plt`) digunakan untuk membuat berbagai jenis grafik. Setelah itu, fungsi `pd.read_csv` digunakan untuk membaca dataset mentah yaitu `data_lengkap.csv` dan hasilnya disimpan dalam variabel `da`.

b. Pra Proses Data

Tahapan pertama yang dilakukan adalah *data cleaning* untuk menangani *missing value* yang terdapat pada dataset awal dengan menghapus setiap baris yang mengandung nilai yang hilang.

```
# Check for missing values and data types
da.isnull().sum()
# Menemukan baris yang memiliki setidaknya satu nilai null
null_data = da[da.isnull().any(axis=1)]
# Menampilkan baris yang memiliki nilai null
null_data

# Menggunakan heatmap untuk visualisasi nilai null
plt.figure(figsize=(10, 6))
sns.heatmap(da.isnull(), cbar=False, cmap='viridis', yticklabels=False)
```

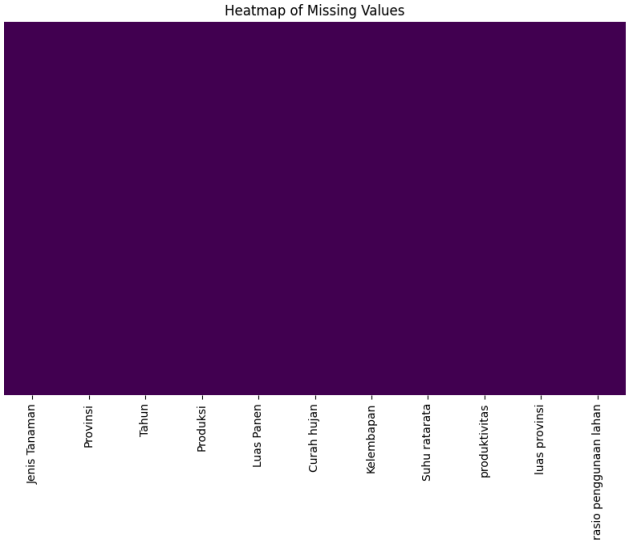
```
plt.title('Heatmap of Missing Values')
plt.show()

data_clean = da.dropna()
data_clean.to_csv('data_cleaning.csv', index=False)

from google.colab import files

# Mengunduh file ke komputer lokal
files.download('data_cleaning.csv')
```

Kode program di atas melakukan pengecekan keberadaan nilai yang hilang di setiap kolom dan menhitung jumlahnya. Baris-baris yang memiliki setidaknya satu nilai null diidentifikasi dan baris-baris yang memiliki nilai null dihapus. Kemudian data cleaning di simpan ke file (`'data_cleaning.csv'`), sehingga menghasilkan perbandingan sebagai berikut :

```
Jenis Tanaman              0
Provinsi                   0
Tahun                      0
Produksi                 112
Luas Panen               112
Curah hujan              125
Kelembapan               260
Suhu ratarata            115
produktivitas            112
luas provinsi              0
rasio penggunaan lahan   112
dtype: int64
```

Gambar 2 Hasil output sebelum *data cleaning*

```
Jenis Tanaman              0
Provinsi                   0
Tahun                      0
Produksi                   0
Luas Panen                 0
Curah hujan                0
Kelembapan                 0
Suhu ratarata              0
produktivitas              0
luas provinsi              0
rasio penggunaan lahan     0
dtype: int64
```



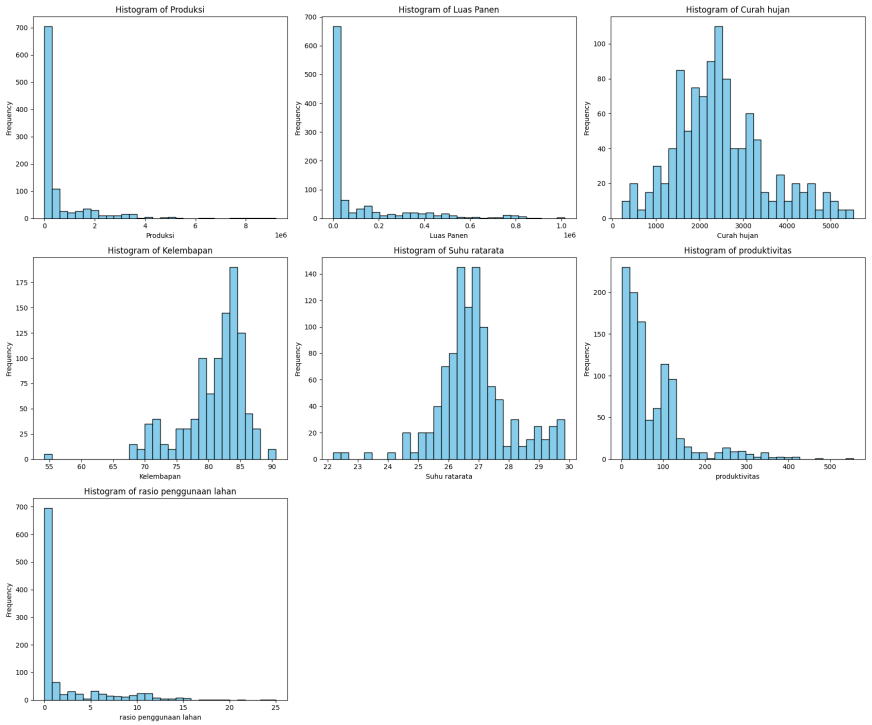
Gambar 3 Hasil output sesudah *data cleaning*

c. Eksplorasi Data

Hal ini dilakukan untuk menganalisis dan memahami karakteristik data yang melibatkan perhitungan statistik deskriptif, hubungan antar variabel dan visualisasi data.

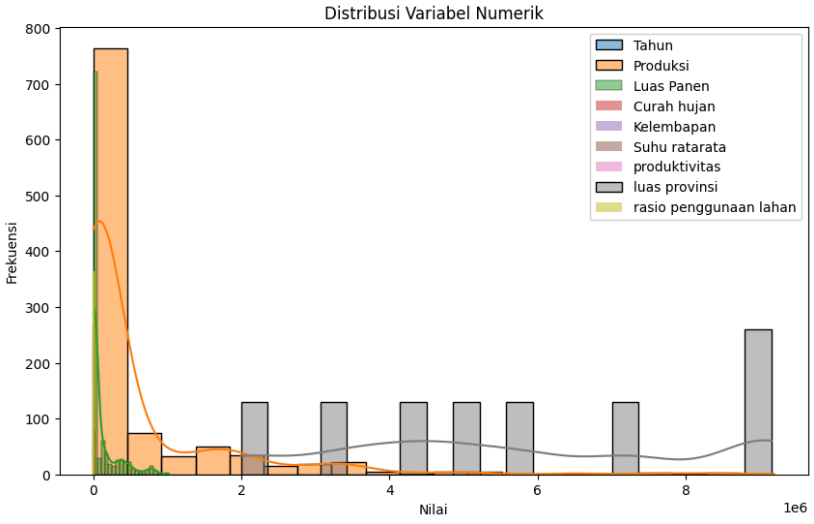
```
display(data.head())

# Statistik deskriptif untuk variabel numerik
data_numerik.describe()

# visualisasi
# Set up the figure and axes for a grid of histograms
fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(18, 15))
# Flatten the axes array for easier iteration
axes = axes.ravel()
# Define columns for histograms
columns = ['Produksi', 'Luas Panen', 'Curah hujan', 'Kelembapan', 'Suhu
ratarata', 'produktivitas',
           'rasio penggunaan lahan']
# Create histograms for each column
for i, col in enumerate(columns):
    axes[i].hist(data[col], bins=30, color='skyblue', edgecolor='black')
    axes[i].set_title(f'Histogram of {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Frequency')
# Hide any unused axes
for j in range(i+1, len(axes)):
    axes[j].set_visible(False)
# Adjust layout to prevent overlap
plt.tight_layout()
# Show the plot
plt.show()

# Visualisasi korelasi antara variabel numerik sebagai heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(data_numerik.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Korelaaasi antara Variabel Numerik')
plt.show()

#reading the database
fig = px.bar(data, x='Jenis Tanaman', y='Produksi', color='Provinsi')
#showing the plot
fig.show()

#reading the database
fig = px.bar(data, x='Provinsi', y='Produksi', color='Jenis Tanaman')
#showing the plot
fig.show()
```

```
#reading the database
fig = px.line(data, y='Produksi', color='Provinsi')

# Menonaktifkan garis grid pada sumbu x dan sumbu y
fig.update_xaxes(showgrid=False)
fig.update_yaxes(showgrid=False)

# Menambahkan area berwarna untuk setiap jenis tanaman
fig.add_vrect(x0=1, x1=208, fillcolor="green", opacity=0.10,
line_width=0)
fig.add_vrect(x0=209, x1=416, fillcolor="yellow", opacity=0.10,
line_width=0)
fig.add_vrect(x0=417, x1=624, fillcolor="red", opacity=0.10,
line_width=0)
fig.add_vrect(x0=625, x1=832, fillcolor="blue", opacity=0.10,
line_width=0)
fig.add_vrect(x0=833, x1=1040, fillcolor="purple", opacity=0.10,
line_width=0)
# Menambahkan anotasi
fig.add_annotation(x=104, y=max(data['Produksi']), text="Padi",
showarrow=False)
fig.add_annotation(x=312, y=max(data['Produksi']), text="Jagung",
showarrow=False)
fig.add_annotation(x=520, y=max(data['Produksi']), text="Kedelai",
showarrow=False)
fig.add_annotation(x=728, y=max(data['Produksi']), text="Ubi Kayu",
showarrow=False)
fig.add_annotation(x=936, y=max(data['Produksi']), text="Ubi Jalar",
showarrow=False)
#showing the plot
fig.show()
```



Gambar 4 Visualisasi histogram

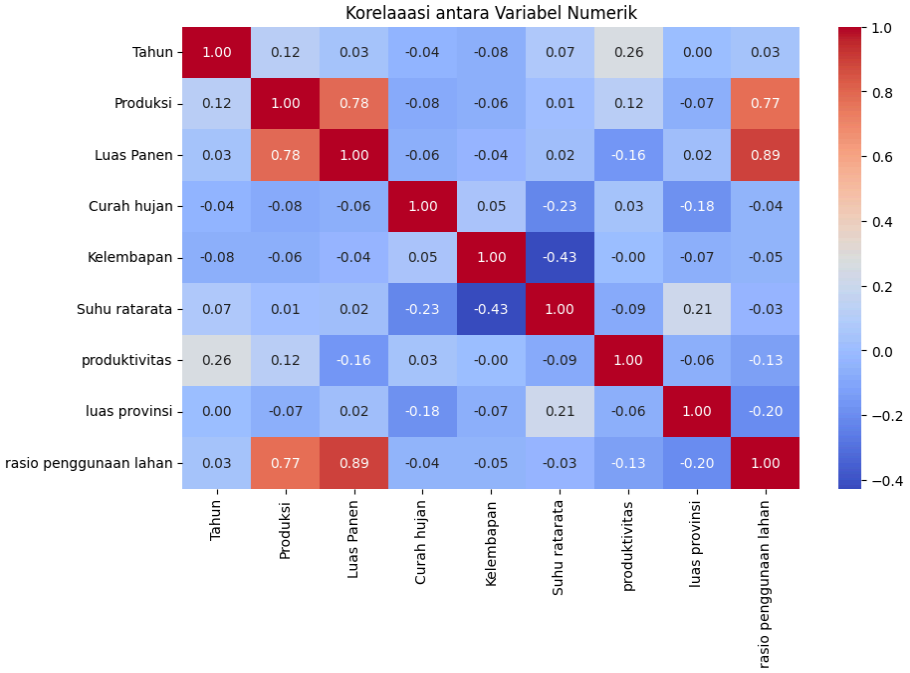
Beberapa histogram yang menggambarkan distribusi variabel numerik dalam dataset. Histogram produksi dan luas panen menunjukkan sebagian besar data terkonsentrasi pada nilai yang sangat rendah, dengan sedikit data pada rentang yang lebih tinggi. Distribusi curah hujan terlihat lebih normal dengan puncak di sekitar 3000 mm. Histogram yang menunjukkan kelembapan dan suhu rata-rata keduanya menunjukkan distribusi yang mendekati normal, dengan kelembapan berpusat di sekitar 80% dan suhu rata-rata di sekitar 27 derajat Celcius. Distribusi produktivitas mirip dengan produksi, di mana sebagian besar data berada pada nilai yang sangat rendah. Rasio penggunaan lahan juga menunjukkan pola yang sama dengan sebagian besar data terkonsentrasi pada nilai yang rendah.



Gambar 5 Visualisasi distribusi variabel numerik

Gambar di atas menunjukkan distribusi variabel numerik dalam dataset. Sumbu horizontal (x) mewakili nilai dari berbagai variabel dan sumbu vertikal (y)

menunjukkan frekuensi kemunculan nilai-nilai tersebut. Setiap warna pada histogram mewakili variabel yang berbeda, seperti Tahun, Produksi, Luas Panen, Curah Hujan, Kelembapan, Suhu Rata-rata, Produktivitas, Luas Provinsi, dan Rasio Penggunaan Lahan. Terlihat bahwa sebagian besar nilai terdistribusi pada kisaran yang lebih rendah dengan beberapa variabel menunjukkan puncak frekuensi yang signifikan di dekat nilai nol.



Gambar 6 Korelasi antara variabel numerik

Heatmap di atas menampilkan korelasi antara beberapa variabel numerik dalam dataset. Sumbu X dan Y mewakili variabel seperti Tahun, Produksi, Luas Panen, Curah Hujan, Kelembapan, Suhu Rata-rata, Produktivitas, Luas Provinsi, dan Rasio Penggunaan Lahan. Warna yang ada di atas menunjukkan nilai korelasi seperti, merah tua menunjukkan korelasi positif sempurna (nilai 1), biru tua menunjukkan korelasi negatif sempurna (nilai -1), dan putih menunjukkan tidak ada korelasi (nilai 0). Beberapa korelasi penting yang terlihat adalah korelasi positif tinggi antara Produksi dan Luas Panen (0.78) serta antara Produksi dan Rasio Penggunaan Lahan (0.77). Korelasi negatif sedang antara Suhu Rata-rata dan Kelembapan (-0.43).

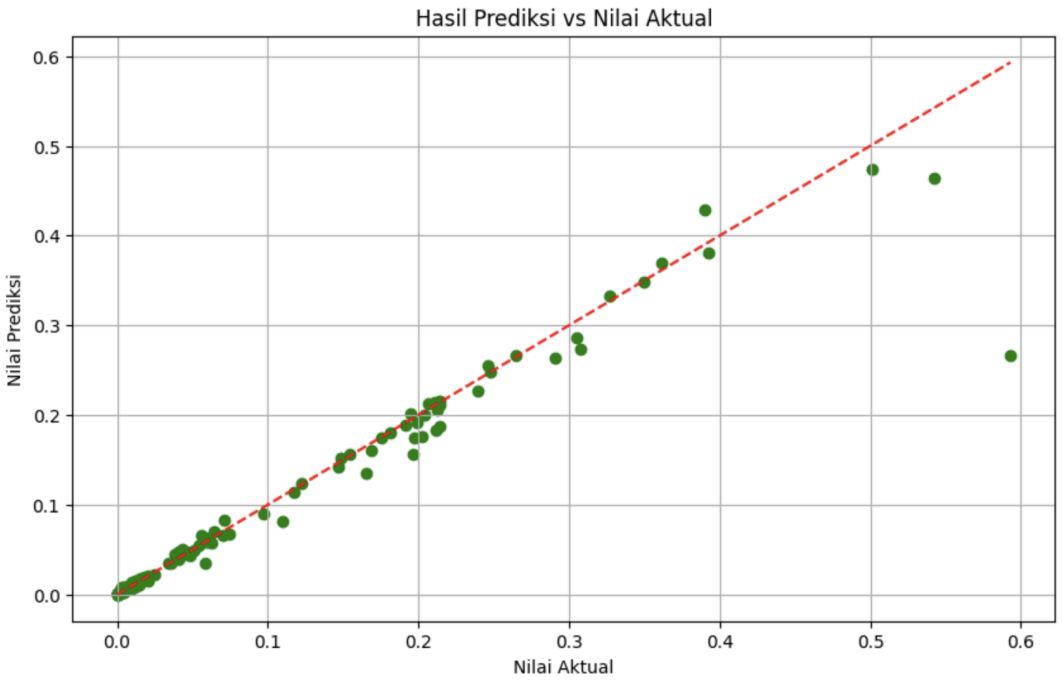
d. Pemodelan Data

Dalam penelitian ini, kami menggunakan model Random Forest Regression (RFR) dalam memprediksi hasil panen tanaman pangan. Random Forest Regression merupakan metode penggabungan (bagging) dari banyak Decision Tree Regresi untuk memprediksi nilai kontinu (Briem et. al. 2001). Tujuan dari penggunaan algoritma ini pada penelitian adalah untuk meningkatkan akurasi prediksi dan mengurangi overfitting. Metrik evaluasi yang digunakan antara ialah Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared Score ($R^2\ Score$).

Tabel 1 Hasil Perhitungan Metrik Evaluasi

Metrik Evaluasi	Nilai
MSE	0.0006031120039796789
MAE	0.005283292395922582
RMSE	0.024558338787053144
$R^2\ Score$	0.9457220340632976

Tahapan pertama dalam melakukan pemodelan data adalah dengan memisahkan fitur (X) dan label (Y) terlebih dahulu, kemudian membagi data menjadi data latih dan data uji. Dalam penelitian ini kami menggunakan 20% dari data sebagai data uji dan 80% dari data sebagai data latih, dikarenakan menurut beberapa sumber bahwa dataset besar sudah sangat cukup dengan menggunakan 10-20% data saja sebagai data latih. Dan untuk memastikan bahwa pembagian data selalu sama setiap kali kode dijalankan, untuk memastikan reproducibility atau konsistensi hasil kami menggunakan '`random_state = 42`'. Setelah itu masuk pada tahap evaluasi model ('`y_pred = model.predict(X_test)`') dan melakukan penghitungan pada MSE.MAE, RMSE serta $R^2\ Score$. Setelah dilakukan uji coba data uji dengan menggunakan RFR didapatkanlah nilai hasil metrik evaluasi tersebut yang dapat dilihat pada **Tabel 1**. MSE digunakan untuk mengukur rata-rata kuadrat dari kesalahan prediksi yang memberikan gambaran seberapa besar kesalahan prediksi secara umum, dan RMSE merupakan akar kuadrat dari MSE yang memberikan gambaran kesalahan dalam satuan yang sama dengan data asli. Sehingga nilai MSE dan RMSE yang semakin besar akan menunjukkan tingkat kesalahan kuadrat yang sangat besar jaraknya diantara nilai aktual dan nilai prediksi. Dalam kasus ini nilai MSE menunjukkan bahwa rata-rata kesalahan kuadrat yang terjadi sangat kecil, yang mengindikasikan bahwa prediksi model cukup dekat dengan nilai sebenarnya, serta niali RMSE yang cukup rendah menunjukkan bahwa model ini memiliki akurasi yang baik. Sedangkan MAE adalah pengukuran rata-rata kesalahan absolut antara nilai aktual dan nilai prediksi. Hal ini menunjukkan seberapa besar kesalahan yang dihasilkan oleh model dalam satuan yang sama dengan target prediksi, dengan nilai MAE pada **Tabel 1**, yang menunjukkan bahwa kesalahan absolut rata-rata prediksi model adalah sekitar 0.0053, yang menunjukkan bahwa tingkat kesalahan relatif kecil. Koefisien determinasi atau $R^2\ Score$ berfungsi untuk mengukur proporsi variabilitas dalam data target. Nilai $R^2\ Score$ berada pada rentang 0 dan 1, jika mendekati nilai 1 maka nilai keakuratan model dapat dikatakan tinggi. Dalam kasus ini, pada Tabel 1 terlihat bahwa nilai $R^2\ Score$ adalah 0.9457220340632976, ini menunjukkan bahwa model dapat menjelaskan 94,57% variasi dalam data hasil panen, yang mengindikasikan kecocokan yang sangat baik antara model dengan data. Visualisasi korelasi antara nilai aktual dan nilai prediksi model RFR ditunjukkan pada Gambar 4.



Gambar 7 Visualisasi Nilai Prediksi vs Nilai Aktual

Gambar 4 merupakan scatter plot yang menunjukkan hubungan antara nilai prediksi model RFR dnegan nilai aktual dari data uji. Sumbu x menggambarkan nilai aktual (‘`y_test`’) dan sumbu y menggambarkan nilai prediksi (‘`y_pred`’). Setiap scatter point berwarna hijau mewakili 1 pengamatan dari data uji. Jika model sempurna, maka semua titik akan berada tepat di garis diagonal merah putus-putus. Garis merah putus-putus berfungsi sebagai referensi untuk melihat seberapa dekat prediksi model dengan nilai aktualnya. Semakin dekat scatter points dengan garis putus-putus merah tersebut maka menunjukkan prediksi yang sangat akurat, begitupun sebaliknya. Dan untuk ditribusi scatter points yang lebih luas dari garis putus-putus merah menunjukkan adanya kesalahan prediksi yang dapat dievaluasi lebih lanjut dengan evaluasi metrik.

KESIMPULAN DAN SARAN

Berdasarkan Nilai-nilai metrik evaluasi didapatkan bahwa model Random Forest Regression yang diterapkan memiliki performa yang sangat baik dengan akurasi 94,57% dalam memprediksi hasil panen tanaman pangan di Sumatera. Nilai MSE, MAE, dan RMSE yang rendah mengindikasikan bahwa kesalahan prediksi yang dihasilkan oleh model sangat kecil. Selain itu, nilai R^2 *Score* yang tinggi menunjukkan bahwa model ini sangat mampu menjelaskan variasi dalam data hasil panen, yang membuatnya menjadi alat prediksi yang efektif. Sebagai saran untuk penelitian berikutnya, jika pada penelitian kali ini kami menggunakan normalisasi sehingga hasil dari metrik evaluasinya berada pada rentang 0 - 1, namun sebenarnya untuk kasus ini tidak diperlukan pra proses normalisasi dikarenakan tidak adanya clustering. Dan diharapkan untuk penelitian lebih lanjut dapat difokuskan pada pengujian model dengan data yang lebih besar dan bervariasi, serta eksplorasi penggunaan fitur tambahan yang mungkin dapat meningkatkan akurasi prediksi lebih lanjut.

DAFTAR PUSTAKA

Bab ini berupa suatu daftar dari semua pustaka yang diacu secara langsung di dalam tubuh tulisan. Contoh penulisan rujukan:

Arinal V, Azhari M. 2023. Penerapan regresi linear untuk prediksi harga beras di indonesia. 5(1):341-346.doi:10.55338/saintek.v5i1.1417

Briem GJ, Benediktsson JA, Sveinsson JR. 2001. Boosting, bagging, and consensus based classification of multisource remote sensing data. *Multiple Classifier Systems: Second International Workshop, MCS 2001 Cambridge, UK, July 2–4, 2001 Proceedings 2*, 279–288.

Fitri E. 2023. Analisis perbandingan metode regresi linier, random forest regression dan gradient boosted trees regression method untuk prediksi harga rumah. *J. Appl. Comput. Sci. Technol.* 4(1):58–64.doi:10.52158/jacost.v4i1.491.

Fitri E, Nugraha SN. 2024. Optimasi kinerja linear regression, random forest regression dan multilayer perceptron pada prediksi hasil panen. *INTI Nusa Mandiri*. 18(2):210–217.doi:10.33480/inti.v18i2.5269.

Gultom FRP, Harianto, Astutik S. 2017. Analisis regresi komponen utama untuk mengatasi multikolinieritas pada kasus kemiskinan di provinsi sumatra utara. *Prosiding Seminar Nasional Matematika dan Pembelajarannya.*

Herlina N, Prasetyorini A. 2020. Pengaruh perubahan iklim pada musim tanam dan produktivitas jagung (zea mays l.) Di kabupaten malang. *J. Ilmu Pertan. Indones.* 25(1):118–128.doi:10.18343/jipi.25.1.118.

Hutahaean J, Mulyani N, Irawati N, Azhar Z, Putri LU. 2024. Analisis prediksi tingkat depresi pada siswa dengan pendekatan regresi linier. *Jurnal Informatika dan Teknologi Informasi*. 2(3):243-252.doi:10.56854/jt.v2i3.330

Isbah U, Iyan RY. 2016. Analisis peran sektor pertanian dalam perekonomian dan kesempatan kerja di Provinsi Riau. [diunduh 2024 Mei 24].

Marutho D. 2019. Perbandingan metode naive bayes, knn, decision tree pada laporan water level jakarta. *INFOKAM*. 15(2):90-97.doi:10.53845/infokam.v15i2.175

Musababa MA. 2023. Implementasi algoritma linear regression untuk prediksi produksi tanaman padi di kabupaten grobogan. *Data Sciences Indonesia (DSI)*. 3(2):68-78. doi:10.47709/dsi.v3i2.3118

Nuraisah G, Kusumo RAB. 2019. Dampak perubahan iklim terhadap usahatani padi di desa wanguk kecamatan anjatan kabupaten indramayu. *Mimb. Agribisnis J. Pemikir. Masy. Ilm. Berwawasan Agribisnis*. 5(1):60–71.

Saadah S, Salsabila H. 2021. Prediksi harga bitcoin menggunakan metode random forest. *Jurnal Politeknik Caltex Riau*. 7(1):24-32.doi:10.35143/jkt.v7i1.4618

Satria A, Badri RM, Safitri I. 2023. Prediksi hasil panen tanaman pangan sumatera dengan metode machine learning. *Digit. Transform. Technol.* 3(2):389–398.doi:10.47709/digitech.v3i2.2852.

Sautomo S, Pardede HF. 2021. Prediksi belanja pemerintah indonesia menggunakan long-short-term-memory (LSTM). Jurnal Resti. 5(1):99-106:doi:10.29207/resti.v5i1.2815

Segal MR. 2004. Machine Learning Benchmark and Random Forest Regression.

Sumaryanto nFN. 2012. Strategi peningkatan kapasitas adaptasi petani tanaman pangan menghadapi perubahan iklim. *Forum Penelit. Agro Ekon.* 30(2):73–89.

Trianggana DA. 2020. Peramalan jumlah siswa-siswi melalui pendekatan metode regresi linear. Jurnal Media Infotama. 16(2):115-120.doi:10.37676/jmi.v16i2.1149

Widhiarso W. 2010. Berkenalan dengan Analisis Mediasi : Regresi dengan Melibatkan Variabel Mediator (Bagian Pertama). [diunduh 2024 Mei 30].

Zainal NK. 2024. Prediksi harga real estate menggunakan metode regresi linear berbasis machine learning. *Journal of Artificial Intelligence Application (JAIA)*. 1(1):19-27.