

Analisis Produktifitas Padi di Pulau Sumatera menggunakan *Exploratory Data Analysis (EDA)*

Rini Mayasari^{1*}, Bagja Nugraha², Ayu Ratna Juwita³, Nono Heryana⁴

¹ Program Studi Informatika, Universitas Singaperbangsa Karawang

^{2,4} Program Studi Sistem Informasi, Universitas Singaperbangsa Karawang

³ Program Studi Informatika, Universitas Buana Perjuangan Karawang

*Corresponding author's email: rini.mayasari@staff.unsika.ac.id

ABSTRAK

Penelitian ini bertujuan untuk menganalisis produktivitas padi di Pulau Sumatera menggunakan *Exploratory Data Analysis (EDA)*. Padi merupakan tanaman pokok yang sangat penting dalam ketahanan pangan di Indonesia, dan Pulau Sumatera memiliki kontribusi signifikan terhadap produksi padi nasional. Namun, faktor-faktor lingkungan dan manajemen pertanian yang memengaruhi produktivitas padi di Pulau Sumatera masih belum sepenuhnya dipahami. Metode EDA digunakan untuk menjelajahi hubungan antara variabel-variabel seperti curah hujan, suhu, kelembapan, luas panen, dan produksi padi di Pulau Sumatera. Melalui analisis deskriptif, visualisasi data, dan eksplorasi hubungan antar variabel, penelitian ini bertujuan untuk memberikan wawasan mendalam tentang faktor-faktor yang mempengaruhi produktivitas padi di wilayah ini. Data yang digunakan dalam penelitian ini diperoleh dari dataset produksi padi di Pulau Sumatera yang mencakup informasi tentang provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata. Melalui langkah-langkah metodologi yang mencakup pengumpulan data, pembersihan data, analisis deskriptif, dan EDA, hasil penelitian ini menyajikan temuan-temuan penting tentang hubungan antar variabel, tren waktu, serta pola dan tren dalam produktivitas padi di Pulau Sumatera. Penelitian ini diharapkan dapat memberikan wawasan yang berharga bagi para pemangku kepentingan di bidang pertanian dan pemerintah dalam upaya meningkatkan produktivitas padi di Pulau Sumatera. Hasil penelitian ini juga dapat menjadi dasar bagi pengembangan strategi yang lebih efektif dalam meningkatkan hasil panen, mengoptimalkan manajemen pertanian, dan pengambilan keputusan yang lebih baik di sektor pertanian.

Kata kunci: Produktivitas padi, Pulau Sumatera, Exploratory Data Analysis (EDA), analisis deskriptif, visualisasi data.

ABSTRACT

This study aims to analyse rice productivity in Sumatra Island using Exploratory Data Analysis (EDA). Rice is a staple crop that plays a crucial role in food security in Indonesia, and Sumatra Island significantly contributes to national rice production. However, the environmental factors and agricultural management practices that influence rice productivity in Sumatra Island are not fully understood. The EDA technique is employed to investigate the relationships between variables such as rainfall, humidity, temperature, rice production, and harvested area on Sumatra Island. Using data visualization, descriptive analysis, and exploration of variable interactions, this study endeavours to offer comprehensive understandings of the factors that influence rice productivity in this region. The study employs data derived from the rice production dataset in Sumatra Island, which contains information concerning province, production, year, harvested area, rainfall, humidity, and average temperature. By adhering to the steps involved in data collection, data cleaning, descriptive analysis, and EDA methodology, this investigation sheds light on crucial findings concerning variable correlations, temporal trends, and patterns that impact rice productivity on Sumatra Island. This research is predicted to provide important insights to the agricultural industry stakeholders and the government for enhancing rice productivity on Sumatra Island. The discoveries of this study can contribute to creating more effective strategies for enhanced crop yields, optimizing agricultural management processes, and making informed decisions in the agricultural sector.

Keywords: Rice productivity, Sumatra Island, Exploratory Data Analysis (EDA), descriptive analysis, data visualization.

1 Latar Belakang

Padi merupakan tanaman pokok yang memiliki peran penting dalam ketahanan pangan di Indonesia, dan Pulau Sumatera secara signifikan berkontribusi terhadap produksi padi nasional [1]. Namun, faktor-faktor lingkungan dan manajemen pertanian yang memengaruhi produktivitas padi di Pulau Sumatera belum sepenuhnya dipahami [2].

Melalui pendekatan EDA [3], penelitian ini bertujuan untuk memberikan wawasan yang lebih mendalam tentang hubungan antara faktor-faktor seperti curah hujan, suhu, kelembapan, luas panen, dan produksi padi di Pulau Sumatera. EDA akan memungkinkan analisis deskriptif, visualisasi data, dan eksplorasi hubungan antar variabel dalam dataset [4]. Dengan memahami hubungan ini, kita dapat mengidentifikasi faktor-faktor penting yang mempengaruhi produktivitas padi dan memberikan dasar bagi pengembangan strategi yang lebih efektif dalam meningkatkan produktivitas padi di Pulau Sumatera.

Penelitian sebelumnya telah menunjukkan bahwa faktor-faktor lingkungan seperti curah hujan, suhu, dan kelembapan memiliki pengaruh yang signifikan terhadap pertumbuhan dan produktivitas padi [5]. Selain itu, luas panen juga merupakan faktor penting yang dapat mempengaruhi produksi padi [6]. Namun, penelitian yang mendalam tentang faktor-faktor ini di Pulau Sumatera masih terbatas.

Penelitian ini bertujuan untuk melengkapi pengetahuan yang ada dengan menggali lebih dalam tentang faktor-faktor yang mempengaruhi produktivitas padi di Pulau Sumatera melalui pendekatan EDA. Dengan memahami faktor-faktor ini, penelitian ini diharapkan dapat memberikan panduan bagi para pemangku kepentingan dalam pengembangan strategi yang lebih efektif dalam meningkatkan produktivitas padi di Pulau Sumatera.

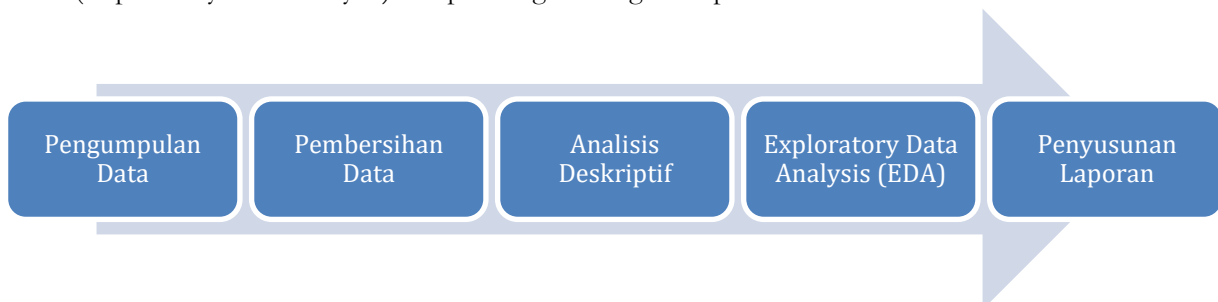
Melalui EDA, peneliti dapat menghasilkan visualisasi data yang informatif [7], seperti scatterplot, line chart, dan heatmap, yang memungkinkan pemahaman yang lebih baik tentang hubungan antar variabel dan tren waktu [8]. Dengan memanfaatkan pendekatan ini, peneliti dapat mengidentifikasi keterkaitan antara luas panen dan produksi padi, mengamati bagaimana faktor lingkungan seperti curah hujan, kelembapan, dan suhu mempengaruhi produktivitas, dan mengevaluasi perubahan dalam produksi seiring waktu.

Penelitian ini bertujuan untuk memberikan wawasan yang berharga kepada para pemangku kepentingan di bidang pertanian dan pemerintah dalam upaya meningkatkan produktivitas padi di Pulau Sumatera. Dengan memahami faktor-faktor yang mempengaruhi produktivitas padi [9], langkah-langkah yang tepat dapat diambil untuk mengoptimalkan manajemen pertanian, penggunaan sumber daya, dan pengambilan keputusan strategis [10] dalam upaya meningkatkan hasil panen.

Dalam penelitian ini, akan dilakukan EDA menggunakan data produksi padi di Pulau Sumatera yang mencakup informasi tentang provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata. Melalui analisis yang mendalam dan interpretasi yang cermat, diharapkan hasil penelitian ini akan memberikan wawasan dan rekomendasi yang berguna dalam upaya meningkatkan produktivitas padi di Pulau Sumatera.

2 Metodologi Penelitian

Metodologi yang digunakan untuk analisis produktivitas padi di Pulau Sumatera menggunakan pendekatan EDA (Exploratory Data Analysis) meliputi langkah-langkah seperti Gambar 1. dibawah ini:



Gambar 1. Tahapan Metodologi Penelitian

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari repositori Kaggle yang berisi dataset produksi padi di Pulau Sumatera [11]. Dataset ini mencakup variabel Provinsi, Tahun, Produksi, Luas Panen, Curah Hujan, Kelembapan, dan Suhu Rata-rata.

2. Pembersihan Data

Pertama-tama, dataset akan diperiksa untuk nilai-nilai yang hilang atau tidak konsisten. Jika ditemukan nilai yang hilang, akan dilakukan imputasi (penggantian nilai yang hilang) atau penghapusan baris, tergantung pada jumlah dan sifat nilai-nilai tersebut. Analisis konsistensi data juga akan dilakukan untuk memeriksa adanya outlier atau nilai yang tidak realistis.

3. Analisis Deskriptif

Setelah dataset dibersihkan, analisis deskriptif akan dilakukan untuk mendapatkan pemahaman dasar tentang data [12]. Ini mencakup perhitungan statistik dasar seperti mean, median, dan modus untuk setiap variabel. Selain itu, distribusi data akan dievaluasi menggunakan plot distribusi dan boxplot.

4. Exploratory Data Analysis (EDA)

EDA akan dilakukan untuk memahami hubungan antara variabel. Ini mencakup pembuatan plot scatter untuk memvisualisasikan hubungan antara variabel, dan mungkin juga mencakup perhitungan korelasi untuk mengukur kekuatan dan arah hubungan antara variabel.

Sebagai bagian dari EDA, analisis tren waktu juga akan dilakukan untuk memahami bagaimana produksi padi dan faktor-faktor lainnya berubah seiring waktu. Ini dapat mencakup plot garis waktu dan mungkin analisis seri waktu lebih lanjut.

5. Penyusunan Laporan

Hasil dari analisis ini akan dirangkum dalam bentuk laporan penelitian, yang mencakup interpretasi dari hasil analisis dan rekomendasi untuk peningkatan produktivitas padi di Pulau Sumatera.

3 Hasil dan Pembahasan

Melalui penggunaan teknik Exploratory Data Analysis (EDA), sejumlah pengamatan kunci telah dibuat terkait dengan dataset produksi padi di Pulau Sumatera.

Tabel 1. Beberapa baris pertama dataframe

| | Provinsi | Tahun | Produksi | Luas Panen | Curah hujan | Kelembapan | Suhu rata-rata |
|---|----------|-------|-----------|------------|-------------|------------|----------------|
| 0 | Aceh | 1993 | 1329536.0 | 323589.0 | 1627.0 | 82.00 | 26.06 |
| 1 | Aceh | 1994 | 1299699.0 | 329041.0 | 1521.0 | 82.12 | 26.92 |
| 2 | Aceh | 1995 | 1382905.0 | 339253.0 | 1476.0 | 82.72 | 26.27 |
| 3 | Aceh | 1996 | 1419128.0 | 348223.0 | 1557.0 | 83.00 | 26.08 |
| 4 | Aceh | 1997 | 1368074.0 | 337561.0 | 1339.0 | 82.46 | 26.31 |

Data yang ditampilkan menunjukkan beberapa baris pertama dari dataset yang digunakan dalam penelitian. Dataset ini mencakup informasi tentang produksi padi di berbagai provinsi di Pulau Sumatera, Indonesia, dari tahun ke tahun.

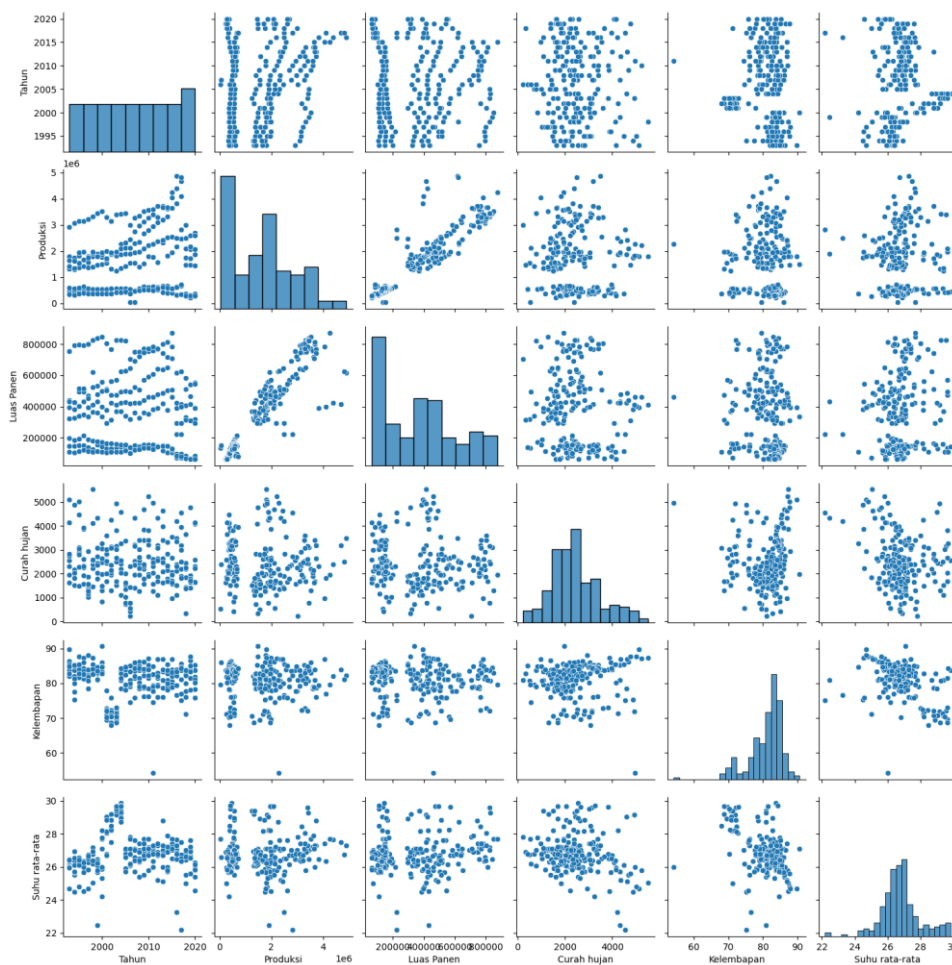
Setelah berhasil mengimpor dataset dan menampilkannya dalam bentuk 'data', tindakan berikutnya yang diambil adalah pengecekan nilai yang hilang. Dari evaluasi yang telah dilakukan, dapat disimpulkan bahwa dataset yang digunakan dalam studi ini tidak memiliki nilai yang hilang. Dengan keyakinan bahwa data tidak memiliki nilai yang hilang, proses berikutnya adalah melaksanakan analisis deskriptif untuk mendapatkan perspektif awal mengenai data yang digunakan.

Tabel 2. Ringkasan Statistik Analisis Deskriptif

| | Produksi | Luas Panen | Curah hujan | Kelembapan | Suhu rata-rata |
|--------------|--------------|---------------|-------------|------------|----------------|
| count | 2.240000e+02 | 224.000000 | 224.000000 | 224.000000 | 224.000000 |
| mean | 1.679701e+06 | 374349.966920 | 2452.490759 | 80.948705 | 26.801964 |
| std | 1.161387e+06 | 232751.161987 | 1031.972625 | 4.878680 | 1.197041 |
| min | 4.293800e+04 | 63142.040000 | 222.500000 | 54.200000 | 22.190000 |
| 25% | 5.488570e+05 | 146919.500000 | 1703.525000 | 78.975000 | 26.177500 |
| 50% | 1.667773e+06 | 373551.500000 | 2315.700000 | 82.375000 | 26.730000 |
| 75% | 2.436851e+06 | 514570.250000 | 3039.700000 | 84.000000 | 27.200000 |
| max | 4.881089e+06 | 872737.000000 | 5522.000000 | 90.600000 | 29.850000 |

Tabel 2 menyajikan ikhtisar statistik yang mencakup rata-rata, median, nilai minimal, dan maksimal untuk setiap kolom dalam DataFrame. Informasi ini memberikan wawasan penting tentang distribusi dan variasi dalam setiap variabel data.

Gambar 2. dibawah ini menampilkan scatterplot dari setiap kombinasi variabel dan histogram untuk setiap variabel tunggal. Hal ini memberikan pandangan mengenai relasi antar variabel serta distribusi masing-masing variabel dalam satu visualisasi.



Gambar 2. Visualisasi data menggunakan pairplot

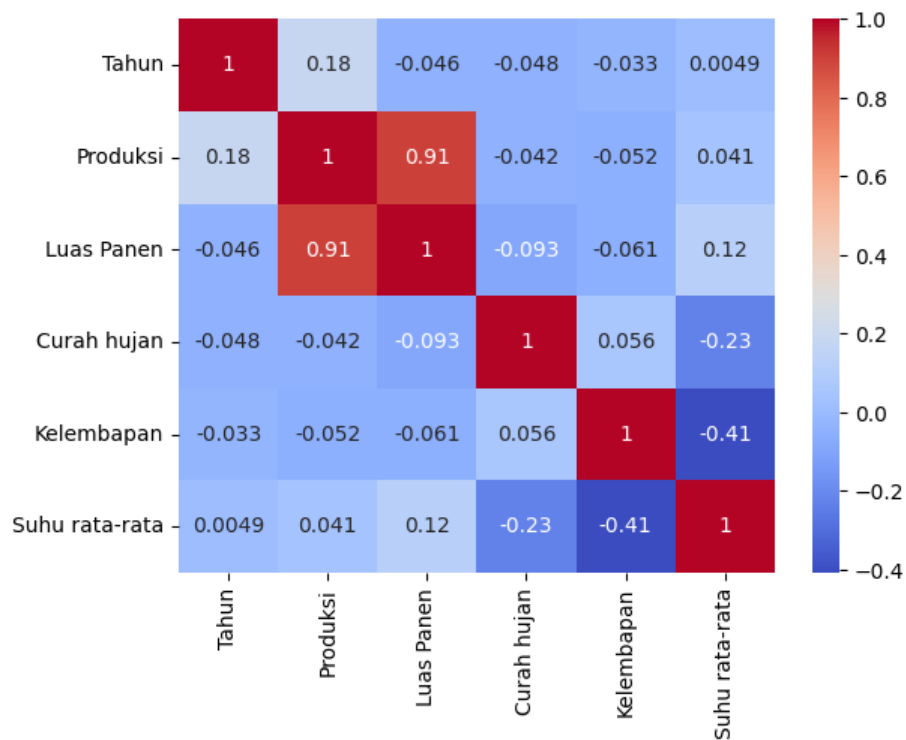
Berikutnya, kita dapat melakukan perhitungan terhadap koefisien korelasi di antara variabel-variabel tersebut untuk memperoleh nilai numerik yang mewakili keterkaitan antara variabel-variabel tersebut.

Tabel 3. koefisien korelasi Pearson antara semua pasangan variabel

| | Tahun | Produksi | Luas Panen | Curah hujan | Kelembapan | Suhu rata-rata |
|----------------|-----------|-----------|------------|-------------|------------|----------------|
| Tahun | 1.000000 | 0.182527 | -0.045951 | -0.047645 | -0.033474 | 0.004923 |
| Produksi | 0.182527 | 1.000000 | 0.905622 | -0.042129 | -0.052316 | 0.041160 |
| Luas Panen | -0.045951 | 0.905622 | 1.000000 | -0.092975 | -0.061121 | 0.115726 |
| Curah hujan | -0.047645 | -0.042129 | -0.092975 | 1.000000 | 0.056466 | -0.228699 |
| Kelembapan | -0.033474 | -0.052316 | -0.061121 | 0.056466 | 1.000000 | -0.407799 |
| Suhu rata-rata | 0.004923 | 0.041160 | 0.115726 | -0.228699 | -0.407799 | 1.000000 |

Tabel 3 diatas menunjukkan hasil perhitungan koefisien korelasi Pearson untuk setiap pasangan variabel dalam dataset. Nilai koefisien korelasi dapat bervariasi dari -1 hingga 1, dimana 1 mengindikasikan adanya hubungan positif yang kuat antara dua variabel, -1 menandakan hubungan negatif yang kuat, dan 0 menunjukkan tidak adanya hubungan antar variabel.

Selanjutnya, untuk memberikan representasi yang lebih intuitif dari hubungan antar variabel, matriks korelasi ditempatkan dalam bentuk heatmap. Heatmap ini memungkinkan visualisasi grafis dari matriks korelasi, memberikan pemahaman visual yang lebih baik tentang bagaimana variabel berkorelasi satu sama lain.



Gambar 3. matriks korelasi menggunakan heatmap

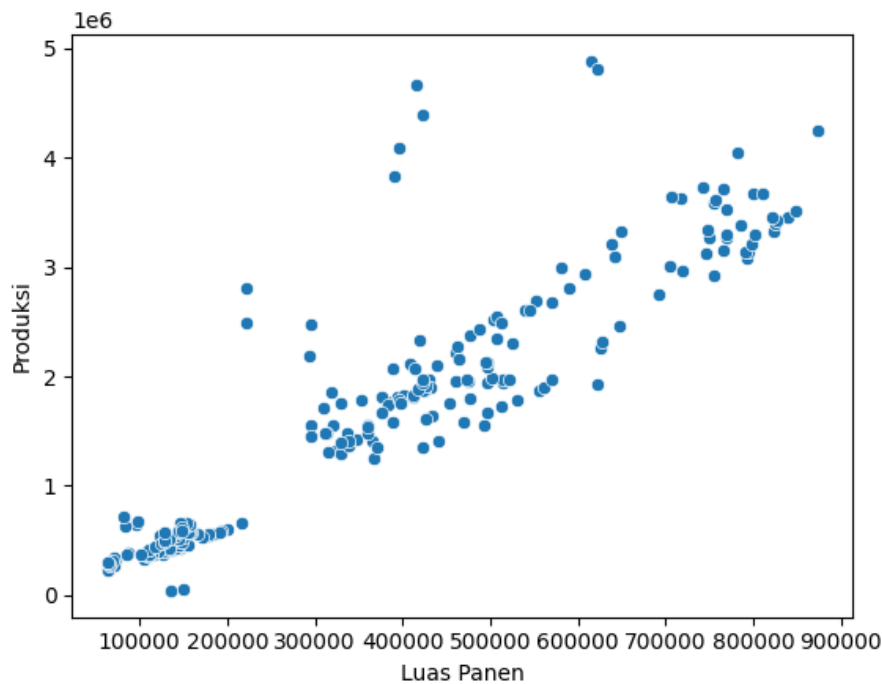
Gambar 3 menampilkan data dalam bentuk representasi visual berupa kotak-kotak berwarna yang disusun dalam suatu matriks. Teknik visualisasi ini dikenal sebagai heatmap. Heatmap merupakan alat yang sangat

efektif untuk menampilkan data dalam format yang mudah dimengerti, dengan menggunakan gradasi warna untuk menunjukkan variasi nilai antar elemen.

Dalam konteks penelitian ini, heatmap digunakan untuk menampilkan matriks korelasi. Matriks korelasi adalah tabel yang menunjukkan koefisien korelasi antara setiap pasangan variabel dalam dataset. Melalui heatmap, kita dapat memahami seberapa erat hubungan antara setiap variabel dengan cara yang lebih intuitif. Warna yang berbeda dalam heatmap mencerminkan kekuatan dan arah korelasi antara variabel: warna yang lebih terang atau gelap (tergantung pada palet warna yang digunakan) menunjukkan korelasi yang lebih kuat, sedangkan warna yang lebih netral menunjukkan korelasi yang lebih lemah.

Melalui penggunaan heatmap, kita dapat dengan cepat mengidentifikasi variabel-variabel yang memiliki korelasi kuat satu sama lain, yang mungkin memerlukan penelitian lebih lanjut untuk memahami hubungan tersebut. Selain itu, heatmap juga memungkinkan kita untuk menemukan pola atau trend yang tidak mungkin terlihat melalui analisis numerik atau statistik saja. Sehingga, heatmap menjadi alat visualisasi yang sangat penting dalam proses analisis data dan penemuan pengetahuan.

Untuk mendapatkan pemahaman yang lebih mendalam tentang data, penting untuk memeriksa lebih rinci keterkaitan antara variabel tertentu. Sebagai contoh, satu aspek menarik yang dapat kita eksplorasi adalah hubungan antara luas area yang ditanam (luas panen) dan total hasil panen (produksi). Alat visualisasi data seperti scatterplot dapat digunakan untuk memvisualisasikan dan memahami hubungan ini dengan lebih baik, seperti yang ditampilkan pada Gambar 4 dibawah ini.

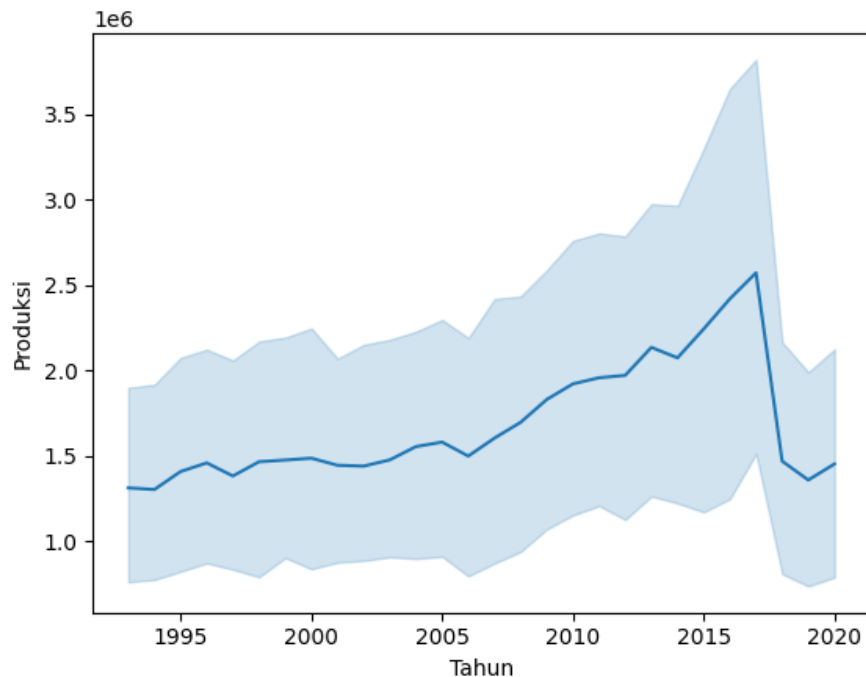


Gambar 4. Visualisasi hubungan antara Luas Panen dan Produksi

Scatterplot adalah alat yang efektif yang memungkinkan kita untuk menunjukkan variasi semantik dalam data kita. Dalam konteks ini, kita menggunakan scatterplot untuk menunjukkan hubungan antara Luas Panen dan Produksi. Melalui scatterplot ini, kita dapat melihat bagaimana setiap titik data (dalam hal ini, setiap pengamatan dari luas panen dan produksi) berinteraksi satu sama lain. Dengan melakukan ini, kita dapat mulai memahami hubungan yang mungkin ada antara luas panen dan hasil produksi padi.

Scatterplot hanya menunjukkan korelasi antara dua variabel dan bukan hubungan sebab-akibat. Oleh karena itu, walaupun kita mungkin melihat pola yang menunjukkan hubungan antara luas panen dan produksi, kita harus berhati-hati untuk tidak langsung menyimpulkan bahwa peningkatan luas panen secara langsung menghasilkan peningkatan produksi. Ada faktor lain yang mungkin mempengaruhi produksi dan perlu dipertimbangkan dalam analisis lebih lanjut.

Kita juga memiliki kemampuan untuk menelusuri perubahan dalam variabel sepanjang berjalannya waktu. Sebagai contoh, kita bisa mengembangkan plot jalur temporal untuk mengevaluasi bagaimana produksi berfluktuasi.



Gambar 5. Plot garis waktu untuk produksi

Sebagaimana ditunjukkan pada Gambar 5, kita menghasilkan plot garis yang mungkin menampakkan variasi semantik. Dalam konteks ini, kita memanfaatkan plot garis untuk menggambarkan evolusi produksi sepanjang periode waktu yang ditentukan. Dengan pendekatan ini, kita dapat lebih memahami dinamika produksi dan mendeteksi pola atau tren yang mungkin terjadi.

4 Kesimpulan

Dapat disimpulkan bahwa analisis produktivitas padi di Pulau Sumatera menggunakan EDA dapat memberikan wawasan yang berharga bagi para pemangku kepentingan di bidang pertanian. Penelitian ini memberikan landasan untuk pengambilan keputusan yang lebih baik dalam manajemen pertanian, penggunaan sumber daya, dan perencanaan strategis untuk meningkatkan produktivitas padi di Pulau Sumatera. Penelitian ini hanya merupakan langkah awal dalam pemahaman produktivitas padi di Pulau Sumatera. Penelitian lanjutan yang melibatkan analisis regresi, pemodelan prediktif, atau studi lebih mendalam dapat memberikan wawasan yang lebih rinci dan mendalam untuk mendukung pengambilan keputusan yang lebih baik di bidang pertanian..

Daftar Pustaka

- [1] N. S. Khodijah, 'Hubungan Antara Perubahan Iklim dan Produksi Tanaman Padi di Lahan Rawa Sumatera Selatan', 2015.
- [2] A. Naumar, Rahmat, and N. Djalir, 'Faktor Penentu Pengelolaan Air Irigasi Untuk Keberlanjutan Ekonomi Pertanian Di Indonesia', *Jurnal Rekayasa*, 2021.
- [3] I. Nurwiana, 'Faktor yang berpengaruh terhadap kinerja sistem irigasi di Wilayah Semi Arid Pulau Timor melalui pendekatan principal component analysis', *Jurnal Irigasi*, 2019.
- [4] J.-D. Fekete, D. Fisher, A. Nandi, and M. Sedlmair, 'Progressive Data Analysis and Visualization (Dagstuhl Seminar 18411)', *Dagstuhl Reports*, vol. 8, pp. 1–40, 2019.

- [5] Y. S. Sari, I. S. Nasution, and S. Syahrul, 'Pengaruh Perubahan Iklim Terhadap Jadwal Tanam Dan Produktivitas Padi Sawah Di Daerah Irigasi (DI.) Krueng Aceh Kabupaten Aceh Besar', *Jurnal Ilmiah Mahasiswa Pertanian*, 2021.
- [6] M. P. R. Rasantaka, M. F. Ashshidiqi, R. Yulianti, Z. Zeinawaqi, and E. S. Widodo, 'Implementasi Regresi Robust untuk Mengetahui Faktor-Faktor yang Mempengaruhi Produksi Padi di Indonesia', *Jurnal Statistika dan Aplikasinya*, 2022.
- [7] R. R. Austin, M. A. Mathiason, and K. A. Monsen, 'Using data visualization to detect patterns in whole-person health data', *Res. Nurs. Health*, vol. 45, no. 4, pp. 466–476, Aug. 2022.
- [8] N. Doshi, S. Gundam, and B. Chaudhury, 'Data Driven and Visualization based Strategization for University Rank Improvement using Decision Trees', *CoRR*, vol. abs/2110.09050, 2021.
- [9] S. A. Handayani, 'Analisis Produksi Dan Pendapatan Usahatani Padidi Desa Pujo Asri Kecamatan Trimurjo Kabupatenlampung Tengah', 2017.
- [10] P. E. Sudjiman and L. S. Sudjiman, 'Analisis Sistem Informasi Manajemen Berbasis Komputer Dalam Proses Pengambilan Keputusan', *TeIKa*, 2020.
- [11] A. Satria, "Dataset tanaman padi sumatera, Indonesia," Kaggle, <https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera> (accessed Jun. 13, 2023).
- [12] Y. Huwaydi and S. F. Persada, 'Analisis Deskriptif Pengguna Go-Pay di Surabaya', *Jurnal Teknik ITS*, vol. 7, 2018.