SMARTICS Journal, Vol.10 No.2 2024. p1-10
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v10i2.11826

Analisis Clustering Produktivitas Padi Daerah-Daerah Di Tiga Provinsi Besar Pulau Jawa Menggunakan Algoritma K-Means Berbasis Python

Sapto Hadi Riono[a,*], Rizdania[b], Henny May Indahsari[c]
[a,b,c]Universitas PGRI Wiranegara, Pasuruan, Indonesia
[*]correspondence email : saptoenator@gmail.com

***Abstract*—Rice plays a crucial role for the Indonesian people as their primary source of energy and carbohydrates. A decline in rice production can impact food availability. Rice, as the main food crop, and its processed products, are crucial in fulfilling more than 70% of the daily food needs for the Indonesian population. One of the methods that can be used to conduct Clustering analysis on rice productivity and rice in three major provinces of Java Island is by using K-Means algorithm. While the tools used in this research is Python. The evaluation result of Sum of Squared Error (SSE) is 3.4452012710511966, Davies-Bouldin Index (DBI) 0.5811451032787581, Silhouette Coefficient Index (SCI) 0.5581692872789441. The use of the K-Means method with Python successfully grouped rice productivity data into three clusters namely high, medium and low. By using fit_predict from the sklearn library, the rice productivity data in the regions of the three major provinces in Java Island were successfully grouped well and gained deep insights and were influenced by various factors such as cultivation techniques, soil quality, and climatic conditions.**

***Index Terms*— Clustering; K-means; Phyton; Paddy.**

***Abstrak*–Bagi masyarakat Indonesia, padi memegang peranan penting sebagai sumber energi dan makanan pokok. Jika produksi beras mengalami penurunan, hal ini dapat berpengaruh terhadap ketersediaan bahan pangan. Padi, sebagai tanaman pangan utama, dan hasil olahannya, sangat penting dalam memenuhi lebih dari 70% kebutuhan pangan sehari-hari bagi penduduk Indonesia. Metode yang dapat digunakan dalam melakukan analisis *Clustering* pada produktivitas padi dan beras di tiga provinsi besar Pulau Jawa adalah dengan menggunakan algoritma *K-Means*. Sedangkan tools yang digunakan pada penelitian ini adalah *Python*. Hasil evaluasi dari *Sum of Squared-Error* (SSE) yaitu 3.4452012710511966, *Davies-Bouldin-Index* (DBI) 0.5811451032787581, *Silhouette-Coefficient Index* (SCI) 0.5581692872789441. Penggunaan metode *K-Means* dengan *Python* berhasil mengelompokkan data produktivitas padi ke 3 *cluster* yaitu tinggi, sedang dan rendah. Dengan menggunakan *fit_predict* dari *library sklearn*, data produktivitas padi di daerah-daerah di tiga provinsi besar di Pulau Jawa berhasil dikelompokkan dengan baik dan mendapatkan wawasan yang mendalam dan dipengaruhi oleh berbagai faktor seperti teknik budidaya, kualitas tanah, dan kondisi iklim.**

***Kata Kunci*— Klustering; K-Means; Phyton; Padi.**

I. PENDAHULUAN

Padi sangat penting bagi masyarakat Indonesia karena berfungsi sebagai sumber utama energi dan karbohidrat bagi mereka[1]. Penurunan produksi beras dapat berpengaruh terhadap ketersediaan pangan, dengan salah satu faktor yang memengaruhi hasil panen padi adalah metode penanaman yang digunakan. Padi, sebagai tanaman pangan utama, dan hasil olahannya, sangat penting dalam memenuhi lebih dari 70% kebutuhan pangan sehari-hari bagi penduduk Indonesia[2]. Kehadiran padi tidak hanya mencakup aspek gizi tetapi juga memiliki dampak langsung pada stabilitas ekonomi dan ketahanan pangan nasional. Pulau Jawa, dengan tiga provinsinya yang signifikan yaitu Jawa Timur, Jawa Tengah, dan Jawa Barat, telah menjadi penopang utama produksi padi di Indonesia.

Namun, meskipun peran sentralnya, produksi padi di Pulau Jawa dihadapkan pada berbagai tantangan yang berpotensi mengganggu stabilitas produksi dan ketahanan pangan. Tantangan-tantangan tersebut meliputi alih fungsi lahan yang mengancam lahan sawah, bencana alam seperti banjir dan kekeringan,

perubahan iklim yang mempengaruhi pola curah hujan dan temperatur, serta keterbatasan teknologi pertanian yang masih dihadapi oleh sebagian besar petani.

Produksi padi di Indonesia didominasi oleh padi sawah, mencakup sekitar 95%, sementara 5% sisanya berasal dari padi ladang atau padi lahan kering. Dalam 25 tahun terakhir, proporsi produksi padi tersebut tetap stabil. Sebagian besar produksi padi nasional berasal dari Pulau Jawa dan Pulau Sumatera, yang secara total menyumbang lebih dari 75% dari produksi padi nasional. Dalam hal ini, sangat penting untuk melakukan analisis menyeluruh terhadap produktivitas padi di tiga provinsi utama Pulau Jawa: Jawa Timur, Jawa Tengah, dan Jawa Barat[3]. Tujuannya adalah untuk membagi daerah berdasarkan hasil panennya menjadi kategori tinggi, sedang, dan rendah. Analisis ini akan memungkinkan untuk mengidentifikasi pola produktivitas yang ada di masing-masing kelompok dan komponen yang mempengaruhinya. Jadi, untuk meningkatkan produksi padi secara berkelanjutan di Pulau Jawa, strategi strategis dapat dibuat.

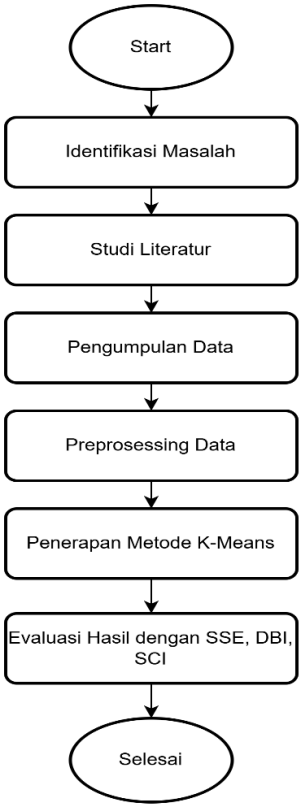
Salah satu metode yang dapat digunakan untuk melakukan analisis *Clustering* pada produktivitas padi dan beras di tiga provinsi besar Pulau Jawa adalah dengan menggunakan algoritma *K-Means*. Sedangkan tools yang digunakan pada penelitian ini adalah *Python*. *Python* dipilih sebagai bahasa pemrograman yang cocok karena popularitasnya dan ketersediaan berbagai pustaka machine learning yang kuat. Dengan menggunakan *Python*, analisis *Clustering* dapat dilakukan dengan fleksibilitas yang tinggi, mulai dari tahap pra-pemrosesan data hingga interpretasi hasil *Clustering*.

Dengan pendekatan *Clustering K-Means*, kita dapat mengelompokkan data produktivitas padi dan beras berdasarkan pola-pola yang muncul. Sebagai contoh, menggunakan algoritma *K-Means*, kita dapat mengidentifikasi kelompok daerah dengan tingkat produktivitas yang rendah secara bersamaan, yang dapat menjadi dasar untuk mengidentifikasi masalah dan mengusulkan solusi yang sesuai. Dengan demikian, analisis *Clustering* dengan *Python* dapat memberikan pemahaman yang lebih dalam tentang faktor-faktor yang memengaruhi produktivitas padi dan beras di tiga provinsi besar Pulau Jawa, serta memberikan landasan untuk pengambilan keputusan yang lebih efektif dalam upaya meningkatkan produksi padi dan beras di wilayah tersebut.

Berdasarkan beberapa penelitian terdahulu yang menerapkan algoritma K-Means seperti Penelitian Teguh Pribadi dan tim berfokus pada penggunaan K-Means untuk mengklasterisasi potensi produksi pertanian di Kabupaten Bojonegoro[4]. Hasil dari klasterisasi menunjukkan pembagian desa ke dalam dua kluster, yaitu kluster baik dan kluster tidak baik. Penggunaan indeks Davies-Bouldin (DBI) sebagai metrik evaluasi juga menunjukkan hasil yang positif dengan nilai DBI = 0.70, yang menampilkan kualitas kluster yang baik. Selanjutnya penelitian oleh Aditya Novita dan tim Penelitian ini bertujuan untuk mengelompokkan provinsi-provinsi menjadi tiga klaster berdasarkan produktivitas komoditas pangan menggunakan algoritma K-Means dan metode Elbow[5]. Selain itu, penelitian yang dilakukan Wahyu Purnomo Priyadi dan tim hasil penelitian menunjukkan bahwa data produksi padi di Jawa Timur dapat dikategorikan ke dalam tiga kelompok utama berdasarkan perbedaan karakteristik produksinya, yaitu rendah, sedang, dan tinggi [6]. Pengujian implementasi web menunjukkan akurasi sebesar 99% dengan tidak adanya selisih antara perhitungan manual K-Means dan perhitungan sistem pada data Kabupaten Pacitan tahun 2018 dan pengujian blackbox sistem juga menunjukkan bahwa hasil input sesuai dengan output dengan persentase 100%. Tidak hanya itu, penelitian yang dilakukan Edi Wahyudin dan tim, Tujuan dari penelitian ini adalah melakukan pengelompokan (clustering) terhadap data produktivitas padi di Provinsi Jawa Barat dengan memanfaatkan algoritma K-Means[7]. Sedangkan pada penelitian ini data yang digunakan adalah produktivitas di tiga pulau jawa dan terdapat 3 kluster yaitu tinggi, sedang dan rendah, kemundian menggunakan evaluasi SCI dan DBI. Pengelompokan data produktivitas padi menggunakan metode K-Means akan mampu membedakan area-area dengan produktivitas tinggi, sedang, dan rendah. Kelompok-kelompok ini dapat memberikan wawasan yang berguna untuk strategi peningkatan produktivitas padi di masa mendatang. Guna menetapkan jumlah kelompok (klaster) yang paling optimal, digunakan operator Cluster Distance Performance yang menilai besaran nilai Davies Bouldin Index (DBI)[7].

II. METODE

Tahapan pelaksanaan penelitian ini berdasarkan pada blok diagram dibawah :



Gambar 1. Tahapan Penelitian

2.1 Identifikasi Masalah

Pada tahap ini, masalah yang diidentifikasi adalah bagaimana mengelompokkan produktivitas padi di daerah-daerah tiga provinsi besar Pulau Jawa (Jawa Timur, Jawa Tengah, dan Jawa Barat) berdasarkan data produktivitasnya, serta mengidentifikasi faktor-faktor yang mempengaruhi produktivitas tersebut.

2.2 Studi Literatur

Studi literatur dilakukan untuk memahami teori-teori terkait produktivitas padi, metode *Clustering K-Means*, dan faktor-faktor yang mempengaruhi produktivitas padi. Literatur yang dikaji meliputi jurnal ilmiah, buku, dan sumber-sumber tepercaya lainnya yang relevan dengan topik penelitian.

2.3 Pengumpulan Data

Data yang dikumpulkan meliputi data produktivitas padi dari tahun 2018 hingga 2022, dari daerah-daerah di tiga provinsi besar Pulau Jawa. Sumber data berasal dari Badan Pusat Statistik (BPS)[8].

2.4 Preprosesing Data

Data yang dikumpulkan akan diproses untuk memastikan kualitas dan konsistensinya. Langkah-langkah *Preprocessing* meliputi pembersihan data (menghilangkan *missing values* dan *outliers*), normalisasi data, dan transformasi data[9]. Pembersihan data melibatkan identifikasi dan penanganan nilai-nilai yang hilang atau *outliers* yang dapat mempengaruhi analisis. Normalisasi data dilakukan untuk memastikan bahwa semua variabel memiliki skala yang sama, sehingga tidak ada variabel yang mendominasi hasil *clustering*. Transformasi data mungkin diperlukan untuk mengubah data ke format yang lebih sesuai untuk analisis

2.5 Penerapan Metode K-Means

Usai melalui tahapan persiapan data (*Preprocessing*), algoritma K-Means diimplementasikan dalam *Jupyter Notebook* untuk melaksanakan pengelompokan (*clustering*) data hasil panen padi berdasarkan kemiripan atributnya [10]. Langkah ini bertujuan untuk mengungkap pola tersembunyi dalam data serta menyajikan perspektif yang lebih mendalam terkait elemen-elemen yang berdampak pada produktivitas padi [11].

2.6 Evaluasi dan Analisis Hasil

Kualitas cluster yang dihasilkan dievaluasi menggunakan *Sum of Squared Error (SSE), Davies-Bouldin Index (DBI), dan Silhouette Coefficient Index (SCI)* serta membantu menentukan seberapa baik data dikelompokkan dan apakah jumlah cluster yang dipilih sudah optimal[10]. SSE mengukur total jarak kuadrat antara setiap titik data dengan *centroid cluster*nya. Semakin kecil nilai SSE, semakin baik kualitas pengelompokan, karena ini menunjukkan bahwa data dalam cluster semakin mendekati *centroid*nya, sehingga *cluster* tersebut lebih kompak dan terdefinisi dengan jelas[11]. DBI berfungsi sebagai tolok ukur kualitas pengelompokan data berdasarkan derajat keterpisahan antar kelompok dan tingkat keutuhan data di dalam kelompok. Nilai DBI yang rendah mengisyaratkan mutu clustering yang lebih tinggi, karena kelompok-kelompok data menjadi lebih berbeda dan seragam. Di sisi lain, SCI menilai tingkat kemiripan data dengan anggota cluster sendiri dibandingkan dengan anggota cluster lain, dengan rentang nilai antara -1 sampai 1. Nilai SCI yang mendekati 1 menandakan hasil clustering yang prima, nilai mendekati 0 mengindikasikan data berada di antara dua kelompok, dan nilai di bawah nol menunjukkan adanya kemungkinan kesalahan dalam penugasan kelompok data. Semakin tinggi nilai SCI, semakin baik kualitas pengelompokan, karena menunjukkan bahwa data dalam *cluster* lebih kompak dan terpisah dengan jelas dari *cluster* lain.

Secara keseluruhan, ketiga metrik evaluasi ini memberikan pandangan yang komprehensif tentang seberapa baik model *clustering* bekerja dalam mengelompokkan data dengan jarak yang terpisah dan homogenitas yang tinggi di dalam cluster.

III. HASIL DAN PEMBAHASAN

3.1 Proses Pengumpulan Data

Perolehan data produktivitas padi dilakukan dengan memanfaatkan informasi yang dipublikasikan oleh Badan Pusat Statistik (BPS) di tiga provinsi utama penghasil padi di Pulau Jawa, yaitu Jawa Timur, Jawa Tengah, dan Jawa Barat, melalui laman daring resmi mereka. Data yang terkumpul meliputi kurun waktu 2018 hingga 2022 serta beragam parameter krusial, termasuk jumlah produksi keseluruhan. Setelah data diunduh, data tersebut akan ditampilkan dalam bentuk tabel terstruktur yang memudahkan dalam visualisasi dan analisis lebih lanjut

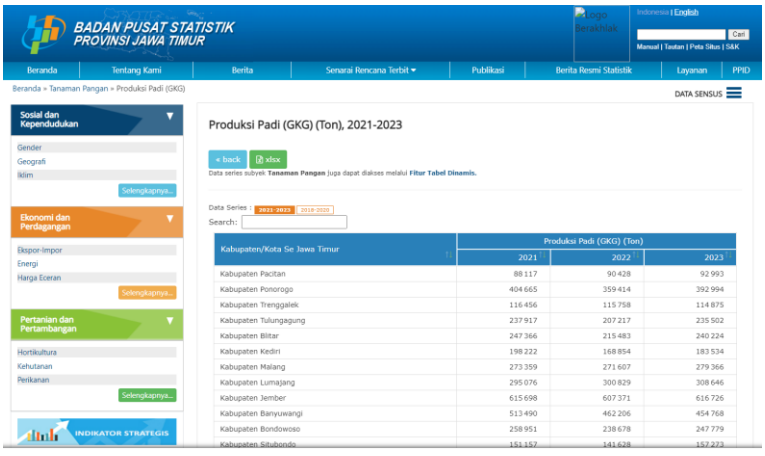
Berikut adalah langkah-langkah untuk proses pengumpulan data:

1: Mengakses Website resmi BPS https://jatim.bps.go.id/, dan memilih kategori statistik pertanian.

2: Menyaring data sesuai dengan provinsi dan periode waktu yang dibutuhkan.

3: Mengunduh dataset dalam Format yang mudah digunakan seperti CSV atau Excel.

4: Mengintegrasikan data dari beragam sumber ke dalam satu file utama bagi setiap provinsi.

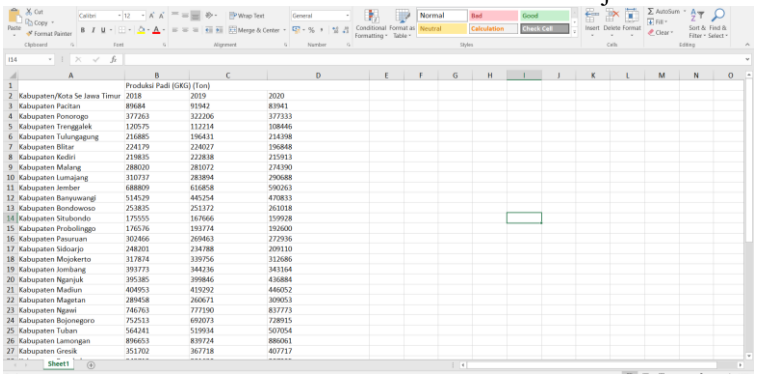


Gambar 2. Website BPS Produksi Padi

3.2 Pembersihan Data

Setelah data diunduh, data tersebut akan ditampilkan dalam bentuk tabel dalam Format (.xlsx) terstruktur yang mencakup semua variabel penting total produksi, dan variabel tahun yang relevan untuk setiap kota dan kabupaten di provinsi Jawa Barat, Jawa Tengah, dan Jawa Timur selama periode 2018 hingga 2022.

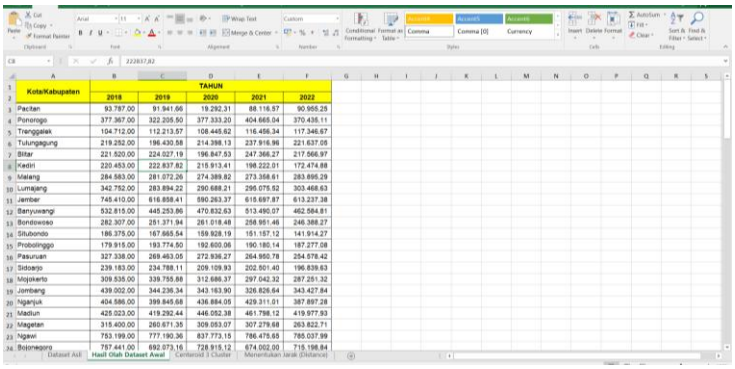
Format tabel akan memudahkan dalam visualisasi dan analisis lebih lanjut.



	A	B	C	D
1		Produksi Padi (GKG) (Ton)		
2	Kabupaten/Kota Se Jawa Timur	2018	2019	2020
3	Kabupaten Pacitan	89684	91942	83941
4	Kabupaten Ponorogo	377263	322206	377333
5	Kabupaten Trenggalek	120575	112214	108446
6	Kabupaten Tulungagung	216885	196431	214398
7	Kabupaten Blitar	224179	224027	196848
8	Kabupaten Kediri	219835	222838	215913
9	Kabupaten Malang	288020	281072	274390
10	Kabupaten Lumajang	310737	283894	290688
11	Kabupaten Jember	688809	616858	590263
12	Kabupaten Banyuwangi	514529	445254	470833
13	Kabupaten Bondowoso	253835	251372	263018
14	Kabupaten Situbondo	175555	167666	159928
15	Kabupaten Probolinggo	176576	193774	192600
16	Kabupaten Pasuruan	302466	269463	272936
17	Kabupaten Sidoarjo	248201	234788	209110
18	Kabupaten Mojokerto	317874	339756	312686
19	Kabupaten Jombang	393773	344236	343164
20	Kabupaten Nganjuk	395385	399846	436884
21	Kabupaten Madiun	404953	415292	446052
22	Kabupaten Magetan	289458	260671	309053
23	Kabupaten Ngawi	746763	777190	837773
24	Kabupaten Bojonegoro	752513	692073	728915
25	Kabupaten Tuban	564241	519934	507054
26	Kabupaten Lamongan	896653	839724	886061
27	Kabupaten Gresik	351702	367718	407717

Gambar 3. Data Setelah di Download dari Website

Dataset siap untuk analisis lebih lanjut dengan outlier yang telah diidentifikasi dan ditangani serta data yang hilang yang telah diisi dengan cara yang tepat dalam Format (.xls) dengan demikian, data yang digunakan dalam analisis terjamin keakuratan, sehingga hasil analisis menjadi lebih dapat diandalkan di tampilkan data nya sebagai berikut:



KotaKabupaten	TAHUN				
	2018	2019	2020	2021	2022
Pacitan	93.787,00	91.941,66	19.282,31	88.118,57	90.966,29
Ponorogo	377.387,00	322.205,50	377.333,20	404.665,64	370.435,11
Trenggalek	104.712,00	112.213,57	108.445,62	116.456,34	117.346,67
Tulungagung	219.252,00	196.430,58	214.398,13	237.916,96	221.637,08
Blitar	221.520,00	224.027,19	196.847,53	247.366,27	217.566,97
Kediri	220.453,00	222.837,82	215.913,41	198.222,01	172.474,88
Malang	284.583,00	281.072,26	274.389,82	273.358,61	283.695,29
Lumajang	342.752,00	283.894,22	290.688,21	295.075,52	303.468,63
Jember	745.410,00	616.858,41	590.263,37	615.697,87	613.237,38
Banyuwangi	532.815,00	445.253,86	470.832,63	513.490,07	462.584,81
Bondowoso	282.307,00	251.371,94	261.018,48	258.951,46	246.388,27
Situbondo	186.375,00	167.665,54	159.928,19	151.157,12	141.914,27
Probolinggo	179.915,00	193.774,50	192.600,06	190.180,14	187.277,08
Pasuruan	327.338,00	269.463,05	272.936,27	264.960,78	254.578,42
Sidoarjo	239.183,00	234.788,11	209.109,93	202.501,40	196.839,63
Mojokerto	309.535,00	339.755,88	312.686,37	297.042,32	287.251,32
Jombang	439.002,00	344.236,34	343.163,90	326.826,64	343.427,84
Nganjuk	404.586,00	399.845,68	436.884,05	429.311,01	387.897,28
Madiun	425.023,00	419.292,44	446.052,38	461.798,12	419.977,93
Magetan	315.400,00	260.671,35	309.053,07	307.279,68	263.822,71
Ngawi	753.199,00	777.190,36	837.773,15	786.475,65	785.037,99
Bojonegoro	757.441,00	692.073,16	728.915,12	674.002,00	719.198,84

Gambar 4. Data Setelah di Tahap Pembersihan Data

Dari hasil identifikasi awal, diperoleh bahwa terdapat sejumlah kabupaten dan kota yang memiliki data lengkap dan valid untuk periode 2018 hingga 2022. Besar sampel akhir adalah sebagai berikut:

- Jawa Timur : 38 kota/kabupaten
- Jawa Tengah : 35 kota/kabupaten
- Jawa Barat : 27 kota/kabupaten

Oleh karena itu, jumlah keseluruhan sampel yang digunakan dalam penelitian ini mencakup 100 kabupaten/kota yang tersebar di tiga provinsi utama di Pulau Jawa. Sampel ini dianggap representatif untuk melakukan analisis produktivitas padi menggunakan metode Clustering K-Means

3.3 Normalisasi Dataset yang akan di Cluster

Setelah memastikan bahwa dataset tidak memiliki nilai kosong, langkah berikutnya adalah melakukan normalisasi data. Proses normalisasi diterapkan dengan metode Min-Max Scaling dari sklearn. Preprocessing, yang menyesuaikan skala variabel ke dalam rentang 0 hingga 1. Normalisasi ini berperan penting dalam memastikan kesetaraan bobot antar variabel dalam analisis Clustering, sehingga hasil Clustering tidak terdistorsi oleh perbedaan skala antar variabel. Normalisasi membantu dalam menyamakan skala data sehingga algoritma K-Means dapat bekerja dengan lebih efektif, proses normalisasi membantu dalam membentuk cluster yang lebih presisi. Tanpa langkah ini, variabel dengan skala lebih besar berisiko mendominasi hasil Clustering, sehingga dapat menimbulkan bias dalam analisis.

Untuk melakukan normalisasi pada *dataset* menggunakan teknik *Min-Max Scaling* dari *Scikit-learn*. Berikut adalah penjelasan langkah demi langkah:

1. Impor *Library MinMaxScaler*:

```Python
from sklearn.preprocessing import MinMaxScaler
```

Kode Program 1 Import Library MinMaxScaler

Kode ini mengimpor `*MinMaxScaler*` dari modul `*sklearn.Preprocessing*`, yang digunakan untuk melakukan normalisasi pada data.

2. Inisialisasi *Scaler*:

```Python
Scaler =MinMaxScaler()
```

Kode Program 2 Inisiasi Scaler

Kode ini membuat instance dari `*MinMaxScaler*`, yang nantinya akan digunakan untuk melakukan normalisasi pada data.

3. Normalisasi Data:

```Python
normalisasi = Scaler.fit_transform(dataset)
```

Kode Program 3 Normalisasi Data

Kode ini melakukan fitting dan *transformasi* pada *dataset*. `fit_trans*For*m()` menghitung minimum dan maksimum untuk setiap kolom dalam *dataset* dan kemudian mengubah nilai-nilai dalam kolom tersebut ke rentang 0 hingga 1 berdasarkan nilai minimum dan maksimum yang telah dihitung.

4. Membuat *DataFrame* Normalisasi:

```Python
dataset_normalisasi = pd.DataFrame(normalisasi, columns=['2018', '2019', '2020', '2021', '2022'])
```

Kode Program 4 DataFrame

Kode ini membuat DataFrame baru dari data yang telah dinormalisasi. Kolom-kolom dalam DataFrame baru ini diberi nama sesuai dengan tahun-tahun yang terdapat dalam *dataset* asli.

5. Menampilkan Data Normalisasi :

```Python
dataset_normalisasi
```

Kode Program 5 Menampilkan DataFrame yg Telah di Normalisasi

Kode ini menampilkan DataFrame yang telah dinormalisasi. DataFrame ini memiliki kolom yang sama seperti *dataset* asli, tetapi dengan nilai yang telah dinormalisasi ke rentang 0 hingga 1.

Normalisasi data bertujuan untuk mengubah skala data ke rentang yang sama, yaitu antara 0 dan 1. Ini penting untuk beberapa alasan:

- Keseimbangan: Memastikan bahwa setiap variabel memberikan kontribusi yang seimbang dalam analisis, terutama dalam algoritma seperti *K-Means Clustering* yang sensitif terhadap skala variabel.
- Konsistensi: Mengurangi efek dari perbedaan skala yang bisa menyebabkan algoritma lebih fokus pada variabel dengan skala yang lebih besar.
- Efisiensi: Mempercepat proses komputasi dengan mengurangi rentang nilai yang harus diproses.

Kode ini mengimpor `*MinMaxScaler*`, menormalisasi data dalam *dataset*, dan menyimpan hasil normalisasi dalam DataFrame baru dengan kolom yang diberi nama sesuai tahun. Normalisasi ini

memastikan bahwa data berada dalam rentang yang sama untuk analisis lebih lanjut, seperti *Clustering*.

	2018	2019	2020	2021	2022
0	0.066124	0.066602	0.014041	0.066652	0.061317
1	0.266608	0.233928	0.276696	0.306561	0.249877
2	0.073848	0.081333	0.079443	0.088131	0.079123
3	0.154825	0.142531	0.157169	0.180185	0.149485
4	0.156428	0.162584	0.144294	0.187346	0.146739
...	...	...	...	...	...
95	0.001558	0.002004	0.001886	0.001616	0.001212
96	0.000071	0.000129	0.000027	0.000055	0.000000
97	0.000097	0.000086	0.000209	0.000205	0.000179
98	0.033007	0.026708	0.026554	0.037096	0.036542
99	0.022167	0.020679	0.020702	0.026869	0.021049

100 rows × 5 columns

Gambar 5. Screenshot Hasil Kode Program Normalisasi Data

Hasil yang ditampilkan adalah DataFrame yang telah dinormalisasi menggunakan metode *Min-Max Scaling*.

Normalisasi data bertujuan untuk mengubah skala data ke rentang yang sama, yaitu antara 0 dan 1. Ini penting untuk beberapa alasan:

- Keseimbangan: Memastikan bahwa setiap variabel memberikan kontribusi yang seimbang dalam analisis, terutama dalam algoritma seperti K-Means Clustering yang sensitif terhadap skala variabel.
- Konsistensi: Mengurangi efek dari perbedaan skala yang bisa menyebabkan algoritma lebih fokus pada variabel dengan skala yang lebih besar.
- Efisiensi: Mempercepat proses komputasi dengan mengurangi rentang nilai yang harus diproses.

	2018	2019	2020	2021	2022
0	0.066124	0.066602	0.014041	0.066652	0.061317
1	0.266608	0.233928	0.276696	0.306561	0.249877
2	0.073848	0.081333	0.079443	0.088131	0.079123
3	0.154825	0.142531	0.157169	0.180185	0.149485
4	0.156428	0.162584	0.144294	0.187346	0.146739
...	...	...	...	...	...
95	0.001558	0.002004	0.001886	0.001616	0.001212
96	0.000071	0.000129	0.000027	0.000055	0.000000
97	0.000097	0.000086	0.000209	0.000205	0.000179
98	0.033007	0.026708	0.026554	0.037096	0.036542
99	0.022167	0.020679	0.020702	0.026869	0.021049

100 rows × 5 columns

Gambar 5. Hasil dari Normalisasi Data

- Baris 0: Nilai-nilai untuk tahun 2018 hingga 2022 adalah 0.066124, 0.066602, 0.014041, 0.066652, dan 0.061317. Ini menunjukkan bahwa data produktivitas pada baris ini relatif rendah dibandingkan dengan maksimum dalam dataset asli.

• Baris 1: Nilai-nilai untuk tahun 2018 hingga 2022 adalah 0.266608, 0.233928, 0.276696, 0.306561, dan 0.249877. Ini menunjukkan bahwa data produktivitas pada baris ini lebih tinggi dibandingkan dengan baris pertama, tetapi tetap dalam rentang 0 hingga 1.
• Baris 98: Nilai-nilai untuk tahun 2018 hingga 2022 adalah 0.033007, 0.026708, 0.026554, 0.037096, dan 0.036542. Ini menunjukkan data produktivitas yang lebih rendah dibandingkan beberapa baris lainnya.

3.4 Menampilkan Jumlah Masing-Masing Cluster Data

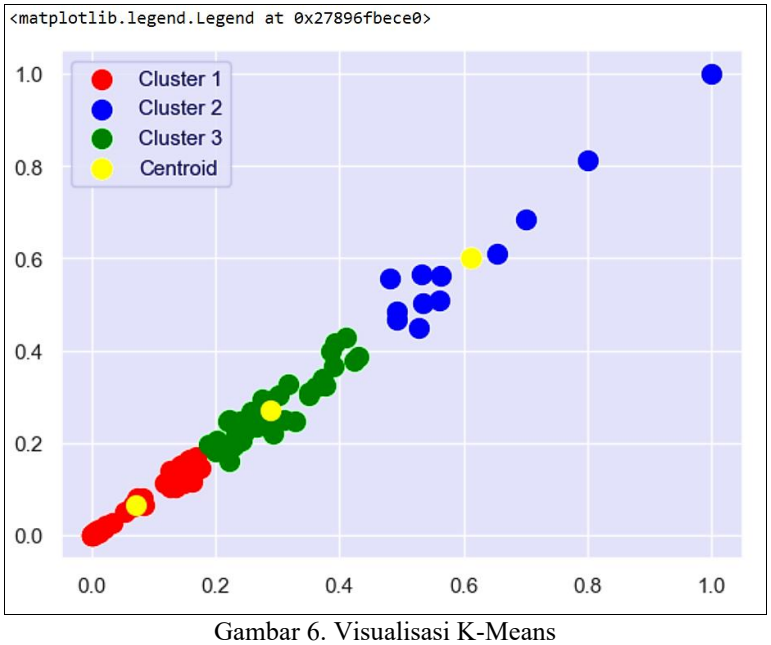
Mengetahui jumlah data dalam setiap cluster membantu dalam memahami distribusi data di antara cluster yang terbentuk. Data ini memiliki peran penting dalam memastikan bahwa hasil pengelompokan tidak menghasilkan cluster yang terlalu kecil maupun terlalu besar, yang bisa menjadi indikasi perlunya penyesuaian parameter Clustering.

Hasil dari perhitungan jumlah data dalam masing-masing cluster berdasarkan output `hasil_cluster` yang telah diberikan sebelumnya.

• Terendah (Cluster 1): Terdapat 51 data yang tergolong dalam cluster ini. Cluster ini memiliki jumlah data terbanyak dibandingkan dengan cluster lainnya.
• Tertinggi (Cluster 2): Terdapat 12 data yang tergolong dalam cluster ini. Cluster ini memiliki jumlah data paling sedikit di antara ketiga cluster.
• Rata-rata/Sedang (Cluster 3): Terdapat 37 data yang tergolong dalam cluster ini. Cluster ini berada di tengah-tengah dari segi jumlah data, tidak terlalu banyak maupun sedikit dibandingkan dengan cluster lainnya.

Total keseluruhan data yang diproses adalah 100, sesuai dengan jumlah total data.
Hasil menunjukkan distribusi data ke dalam tiga cluster setelah dilakukan pemodelan dengan algoritma K-Means. Cluster dengan jumlah data terbanyak adalah cluster 1, diikuti oleh cluster 3, dan cluster 2 memiliki jumlah data paling sedikit.

3.5 Menampilkan Hasil Cluster dalam Bentuk Grafik



Gambar 6. Visualisasi K-Means

Hasil dari *plot* ini adalah *visualisasi* yang memungkinkan untuk melihat bagaimana data terdistribusi ke dalam tiga *cluster* berbeda berdasarkan posisi mereka dalam ruang fitur. *Centroid* ditampilkan sebagai titik-titik kuning yang merupakan representasi dari titik pusat masing-masing *cluster* berdasarkan koordinatnya. Dengan visualisasi ini, dapat dilihat pola atau struktur yang dihasilkan oleh algoritma *K-Means* dalam mengelompokkan data.

3.6 Menampilkan Nilai *Sum of Squared Error* (SSE), *Davies Bouldin Index* (DBI) dan *Silhouette Coefficient Index* (SCI)

- Hasil dari *Sum of Squared Error* (SSE) yang diberikan, yaitu `3.4452012710511966`, mengindikasikan tingkat variabilitas atau penyebaran data dalam *cluster* yang dihasilkan oleh model K-Means. Semakin rendah nilai SSE, semakin baik model *K-Means* dapat memisahkan data menjadi *cluster* yang berbeda, karena cluster tersebut memiliki titik-titik data yang lebih dekat dengan *centroid*nya.
- *Davies Bouldin Index* (DBI) = 0.5811451032787581
  DBI adalah metrik untuk mengevaluasi seberapa baik suatu algoritma *clustering* menghasilkan kelompok data yang padat di dalamnya dan berjauhan antar kelompok [13]. Semakin kecil nilai DBI, semakin optimal hasil *clustering*nya karena cluster menjadi lebih homogen dan distinktif [13]. Dalam kasus ini, nilai DBI = 0.5811 mengindikasikan bahwa model telah membentuk *cluster* dengan tingkat kepadatan internal dan pemisahan eksternal yang memadai, menunjukkan kualitas *clustering* yang cukup baik. Meskipun tidak berada pada nilai yang sangat rendah, hasil ini masih dalam kategori yang cukup baik untuk menggambarkan bahwa *clustering* sudah dapat membedakan kelompok dengan baik.
- *Silhouette Coefficient Index* (SCI) = 0.5581692872789441

SCI merupakan indikator yang menilai tingkat kedekatan data dalam sebuah kelompok dibandingkan dengan data di kelompok lainnya. Nilai SCI antara -1 hingga 1, untuk nilai mendekati 1 menunjukkan *clustering* yang sangat baik, nilai mendekati 0 menunjukkan bahwa data hampir pada batas antara dua *cluster*, dan nilai negatif menunjukkan kemungkinan bahwa data dikelompokkan ke *cluster* yang salah[11]. Dalam hal ini, nilai SCI sebesar 0.5581 menunjukkan *clustering* yang cukup baik. Nilai ini menunjukkan bahwa data dalam *cluster* cukup kompak dan terpisah dengan jelas dari data di cluster lain, namun masih ada ruang untuk peningkatan kualitas *clustering* agar lebih optimal.

IV. KESIMPULAN

1. Penggunaan metode K-Means dengan Python berhasil mengelompokkan data produktivitas padi ke dalam beberapa cluster berdasarkan kesamaan karakteristiknya. Dengan menggunakan *fit_predict* dari *library sklearn*, data produktivitas padi di daerah-daerah di tiga provinsi besar di Pulau Jawa berhasil dikelompokkan dengan baik, menghasilkan *cluster-cluster* yang dapat dianalisis lebih lanjut untuk mendapatkan wawasan yang mendalam.
2. Hasil dari evaluasi kualitas clustering, didapatkan nilai Davies-Bouldin Index (DBI) sebesar 0.5811451032787581 dan Silhouette Coefficient Index (SCI) sebesar 0.5581692872789441. artinya nilai DBI sebesar 0.5811 menunjukkan bahwa cluster yang dihasilkan oleh model cukup baik, dengan tingkat pemisahan antar cluster yang cukup jelas dan kedekatan data yang tinggi di dalam cluster masing-masing sedangkan nilai SCI sebesar 0.5581 menunjukkan bahwa data dalam cluster cukup kompak dan terpisah dengan jelas dari data di cluster lain, namun masih ada ruang untuk peningkatan kualitas clustering agar lebih optimal

DAFTAR PUSTAKA

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
[2] Mergono Adi Ningrat, Carolina Diana Mual, and Yohanis Yan Makabori, "Pertumbuhan dan Hasil Tanaman Padi (Oryza sativa L.) pada Berbagai Sistem Tanam di Kampung Desay, Distrik Prafi, Kabupaten Manokwari," Prosiding Seminar Nasional Pembangunan dan Pendidikan Vokasi Pertanian, vol. 2, no. 1, 2021, doi: 10.47687/snppvp.v2i1.191.
[3] U. Maman, I. Aminudin, and E. Novriana, "Efektifitas Pupuk Bersubsidi Terhadap Peningkatan Produktivitas Padi Sawah," *Jurnal Agribisnis Terpadu*, vol. 14, no. 2, 2021, doi: 10.33512/jat.v14i2.13268.
[4] M. Masganti, A. Susilawati, and N. Yuliani, "Optimasi Pemanfaatan Lahan untuk Peningkatan Produksi Padi di Kalimantan Selatan," *Jurnal Sumberdaya Lahan*, vol. 14, no. 2, 2020, doi: 10.21082/jsdl.v14n2.2020.101-114.
[5] T. Pribadi, R. Irsyada, H. Audytra, and D. A. Fatah, "Implementasi Algoritma K-Means Untuk Klasterisasi Potensi Desa Pada Sektor Produksi Pertanian Di Kabupaten Bojonegoro," *Jurnal Simantec*, vol. 9, no. 1, 2020, doi: 10.21107/simantec.v9i1.9006.
[6] A. Novita, Ii. Ernawati, and N. Chamidah, "Klasterisasi Provinsi Di Indonesia Berdasarkan Produktivitas Komoditas Pangan Menggunakan Algoritma K-Means," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 2022.
[7] W. P. Priyadi, J. Dedy Irawan, and A. Faisol, "Penerapan Data Mining Untuk Clustering Wilayah Produksi Padi Menggunakan Metode K-Means (Studi Kasus : Wilayah Jawa Timur)," 2024.
[8] E. Wahyudin, R. Amir Rudin, and S. Eka Permana, "Penerapan Data Mining Pengelompokan Produktivitas Padi Menggunakan Algoritma K-Means Pada Provinsi Jawa Barat," 2024. [Online]. Available: https://opendata.jabarprov.go.id/id/dataset/produktivit
[9] BPS, "Dataset Produktivitas Padi," BPS. Accessed: Apr. 01, 2024. [Online]. Available: https://searchengine.web.bps.go.id/search?mfd=3500&q=produksi+panen+padi&content=all&page=1&title=0&from=all&to=all&sort=relevansi

[10] Davy Cielen; Arno Meysman, Introducing Data Science: Big data, machine learning, and more, using Python tools. Manning, 2016.
[11] M. S. H. M. S. E. Denny Jolyta, Teknik Evaluasi Cluster Solusi Menggunakan Python Dan Rapidminer. Deepublish, 2021.
[12] M. Syafrullah, "Pengklasteran dan Segmentasi Karakteristik Donatur Sedekah Daring Dengan Teknik Penambangan Data," vol. 4, no. 1, p. 2020.

Sapto Hadi Riono, Meraih gelar Sarjana Komputer pada tahun 2013 dari STMIK Yadika Bangil. Kemudian meraih gelar Magister Komputer dari Universitas Dian Nuswantoro pada tahun 2018. Saat ini Penulis menjadi dosen Program Studi Ilmu Komputer Universitas PGRI Wiranegara Pasuruan.

Rizdania, Meraih gelar Sarjana Teknik pada tahun 2004 dari Universitas Brawijaya Malang. Kemudian meraih gelar Magister Komputer dari Universitas Brawijaya Malang pada tahun 2018. Saat ini Penulis menjadi dosen Program Studi Ilmu Komputer Universitas PGRI Wiranegara Pasuruan

Henny May Indahsari, Mahasiswa Program Studi Ilmu Komputer Fakultas Teknologi dan Sains Universitas PGRI Wiranegara Pasuruan.