Maliki Interdisciplinary Journal (MIJ) eISSN: 3024-8140
Volume 2, Issue 1, 2024 pp. 128-137 http://urj.uin-malang.ac.id/index.php/mij/index

K-Means Clustering untuk pengelompokan daerah penghasil padi di Indonesia berdasarkan luas panen, produksi, dan produktivitas padi tahun 2022

Erika Ayu Prastia Putri
Program Studi Matematika, Universitas Islam Negeri Maulana Malik Ibrahim Malang
e-mail: 200601110092@student.uin-malang.ac.id

Kata Kunci:
kuantitatif; data mining; klustering; metode k-means; tanaman padi

Keywords:
quantitative; data mining; clustering; k-means method; rice plants

A B S T R A K

Tanaman padi merupakan salah satu tanaman pokok sebagai kebutuhan utama masyarakat Indonesia dalam pemenuhan bahan pangan. Sebesar 10,61 juta ha wilayah daratan Indonesia digunakan sebagai lahan pertanian padi tentunya diharapkan oleh pemerintah agar dapat memenuhi kebutuhan pangan seluruh masyarakat Indonesia. Setiap wilayah di Indonesia tentunya memiliki luas panen padi, produksi, dan produktivitas yang berbeda. Terdapat beberapa daerah yang menghasilkan cukup banyak padi dan terdapat pula beberapa daerah yang tidak cukup banyak menghasilkan padi. Beranjak dari masalah ini kita dapat membagi daerah di Indonesia dengan beberapa kategori daerah penghasil padi. Akan dibentuk 3 cluster pada pengelempokan ini, yaitu: (1) Cluster 1 Daerah Penghasil Padi Terbesar, (2) Cluster 2 Daerah Penghasil padi Menengah, (3) Cluster 3 Daerah Penghasil Padi Terendah. Perhitungan cluster dengan algoritma K-Mean menggunakan software Minitab dengan visualisasi berupa 3D Scatterplot.

A B S T R A C T

Rice plants are one of the main crops for fulfilling the food needs of the Indonesian people. Out of the 10.61 million hectares of land in Indonesia, used for agriculture, rice cultivation covers a significant portion, and the government hopes to meet the food needs of the entire population. Each region in Indonesia has different rice harvest areas, production, and productivity levels. Some areas produce a considerable amount of rice, while others do not. Based on this issue, we can divide the regions in Indonesia into several categories of rice-producing areas. This clustering will form three clusters, namely: (1) Cluster 1, the Largest Rice-Producing Region, (2) Cluster 2, the Intermediate Rice-Producing Region, and (3) Cluster 3, the Lowest Rice-Producing Region. The cluster calculation is performed using the K-Mean algorithm with Minitab software, and the visualization is presented in a 3D Scatterplot.

Pendahuluan

Tanaman padi merupakan salah satu tanaman pokok sebagai kebutuhan utama masyarakat Indonesia dalam pemenuhan bahan pangan. Mengingat wilayah Indonesia yang sangat luas dengan luas daratan sebesar 1,91 juta km^2 dan luas wilayah perairan mencapai 6,32 juta km^2.Menurut data dari Badan Pusat Statistik (BPS), sebesar 10,61 juta ha wilayah daratan Indonesia digunakan sebagai lahan pertanian padi di seluruh wilayah Indonesia. Laporan World Population Review mencatat, jumlah penduduk Indonesia mencapai 275,5 juta orang hingga 1 November 2022. Jumlah ini menempatkan

This is an open access article under the CC BY-NC-SA license.
Copyright © 2023 by Author. Published by Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Indonesia berada di peringkat keempat penduduk terbanyak diantar negara G20. Dengan banyaknya penduduk Indoensia yang mencapai 275,5 juta orang tentunya juga berdampak terhadap meningkatnya jumlah kebutuhan pokok yang harus dipenuhi. Salah satu kebutuhan pokok wajib masyarakat Indonesia yaitu beras.

Sebesar 10,61 juta ha wilayah daratan Indonesia digunakan sebagai lahan pertanian padi tentunya diharapkan oleh pemerintah agar dapat memenuhi kebutuhan pangan seluruh masyarakat Indonesia. Namun, sangat disayangkan dengan luasnya daerah lahan pertanian padi tidak mampu mecukupi kebutuhan pangan beras masyarakat Indonesia. Hal tersebut, dapat dilihat dari kegiatan impor beras yang masih dilakukan Indonesia hingga saat ini. Badan Pusat Statistik (BPS) mencatat, Indonesia mengimpor beras sebanyak US$202,04 juta pada 2022. Nilai tersebut naik sebesar 9,92% dibandingkan tahun sebelunya yang sebesar US$183,80 juta. Perlu kita ketaui hal apa saja yang menyebabkan Indonesia masih melakukan impor beras dari negara lain, diantaranya yaitu :

1. Permintaan yang tingggi
2. Kurangnya produktivitas
3. Kerusakan hasil panen
4. Cuaca buruk yan menyebabkan produksi terangu

Dari penyebab tersebut perlu adanya evaluasi agar lahan yang luas untuk penanaman padi dapat diolah dengan optimal dan dapat sepenuhnya memenuhi kebutuhan pangan seluruh masyarakat Indonesia.

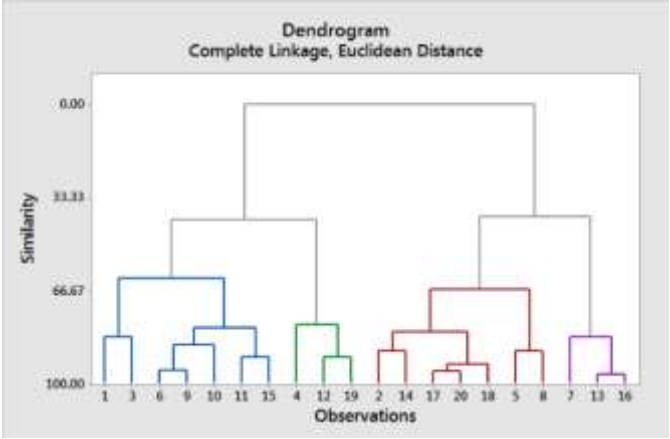
Setiap wilayah di Indonesia tentunya memiliki luas panen padi, produksi, dan produktivitas yang berbeda. Terdapat beberapa daerah yang menghasilkan cukup banyak padi dan terdapat pula beberapa daerah yang tidak cukup banyak menghasilkan padi. Beranjak dari masalah ini kita dapat membagi daerah di Indonesia dengan beberapa kategori daerah penghasil padi. Daerah dengan penghasil padi utama atau tinggi diharap mampu terus mengoptimalkan hal tersebut agar hasil panen padi mereka dapat disebarkan luaskan di seluruh wilayah Indonesia. Kemudian daerah dengan hasil panen padi rendah diharap dapat mulai mengevaluasi apa penyebab hal tersebut terjadi dan dapat memperbaikinya agar produksi padi dapat meningkat. Dengan ini, pengelompokkan daerah penghasil padi serta pengkategorian daerah penghasil padi diperlukan untuk mengoptimalkan produksi padi.

Analisis *Cluster*

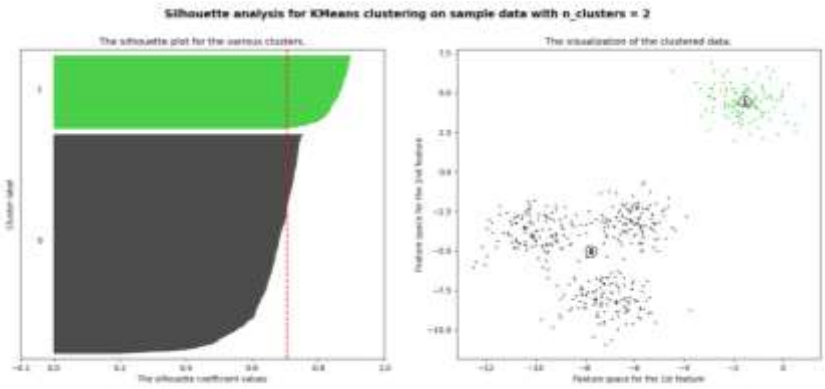
Analisis *cluster* merupakan suatu pengelompokkan objek data yang berdasarkan informasi yang ditemukan dalam data yang menggambarkan objek dan hubungannya. Tujuannya adalah agar objek-objek dalam suatu kelompok serupa (atau terkait) satu sama lain dan berbeda dari (atau tidak terkait) objek-objek dalam kelompok lain (Tan, Steinbach, & Kumar, 2005). Misalnya, pengelompokan dapat dianggap sebagai bentuk klasifikasi yang menciptakan pelabelan objek dengan label kelas (*cluster*). Objek baru yang tidak berlabel diberi label kelas menggunakan model yang dikembangkan dari objek dengan label kelas yang diketahui. Secara logika, *cluster* yang baik adalah *cluster* yang mempunyai: (1) Homogenitas (kesamaan) yang tinggi antar anggota dalam satu

cluster, (2) Heterogenitas (perbedaan) yang tinggi antar *cluster* yang satu dengan *cluster* yang lainnya (Rahman, 2019).

Opsi visualisasi *cluster* meliputi dendrogram dan plot siluet. Dendoram adalah grafik pohon yang digunakan dalam analisis *cluster* untuk menunjukkan hubungan hierarkis antara objek dalam suatu kumpulan data (Rahman, 2019). Dalam analisis *cluster*, dendogram digunakan untuk memvisualisasikan bagaimana objek-objek dikelompokan ke dalam *cluster* yang lebih besar berdasarkan kesaaman mereka. Plot siluet adalah grafik yang digunakan untuk mengevaluasi kualitas klasterisasi dari model *clustering*. Plot siluet menggambarkan nilai siluet untuk setiap objek dalam dataset dan memberikan gambaran tentang seberapa baik objek dikelompokkan ke dalam *cluster* tertentu . Plot siluet dapat membantu dalam menentukan jumlah *cluster* yang optimal dan juga membantu dalam memperbaiki model *clustering* yang tidak baik. Jika nilai siluet tinggi dan sebagian besar objek memiliki nilai siluet positif maka klasterisasi dikatakan baik. Namun, jika nilai siluet rendah dan sebagian objek memiliki nilai siluet negatif maka klasterisasi tersebut buruk dan perlu diperbaiki.



Gambar 1. Dendogram.



Gambar 2. Plot Siluet (*silhouette*)

K-Means Clustering

K-Means Clustering adalah algoritma *unsupervised learning* dalam *machine learning* yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok atau *cluster*

berdasarkan kemiripan fitur antar data (Muller & Guido, 2016). *Unsupervised learning* sendiri adalah salah satu jenis pembelajaran mesin dimana algoritma harus mempelajari pola atau struktur dalam data tanpa memiliki label atau informasi target yang spesifik.

Algoritma K-Means Clustering membagi data menjadi K kelompok, dimana setiap kelompok diwakili oleh pusat *cluster* atau *centroid*. *K-Means Clustering* bekerja dengan cara menghitung jarak antara setiap data dan pusat *cluster* terdekat, lalu memasukkan data tersebut ke dalam kelompok yang sesuai dengan *centroid* terdekat. Selanjutnya, *centroid* dari setiap kelompok dihitung kembali sebagai rata-rata dari data yang termasuk dalam kelompok tersebut. Proses ini diulang hingga tidak ada lagi perubahan dalam penempatan data ke dalam kelompok (Muller & Guido, 2016).

Dalam melakukan *K-Means Clustering* terdapat beberapa hal yang perlu diperhatikan, diantaranya adalah:

1. Jumlah *cluster* : Pemilihan jumlah *cluster* (K) sangat penting dalam *K-Means Clustering*. Jika K terlalu kecil, maka informasi dalam data mungkin akan hilang. Sedangkan jika K terlalu besar, maka hasil clustering akan menjadi terlalu detail dan sulit untuk diinterpretasikan. Oleh karena itu, pemilihan K yang tepat dapat dilakukan dengan metode seperti *Elbow Method* atau *Silhouette Score*.
2. Pemilihan *centroid* awal : Proses *K-Means clustering* sangat sensitif terhadap pemilihan *centroid* awal. Oleh karena itu, pemilihan *centroid* awal yang baik dapat mempengaruhi hasil *clustering* secara signifikan. Beberapa metode yang umum digunakan untuk pemilihan *centroid* awal antara lain *Random Initialization, K-Means++ Initialization*, dan *Fuzzy C-Means Initialization.*
3. Normalisasi data : Pada umumnya, normalisasi data diperlukan dalam *K-Means clustering* karena data yang memiliki skala yang berbeda-beda dapat mempengaruhi hasil *clustering*. Dalam normalisasi data, setiap fitur dalam data dinormalisasi ke rentang nilai yang sama.
4. Konvergensi : *K-Means clustering* merupakan algoritma iteratif dan berhenti saat tidak ada lagi perubahan dalam penempatan data ke dalam *cluster*. Oleh karena itu, perlu memperhatikan kriteria berhentinya iterasi agar tidak terjadi *overfitting* atau *underfitting*.

Metode

Pada *K-Means clustering* terdapat banyak pendekatan untuk membuat *cluster*, diantaranya adalah membuat aturan yang mengelompokan keanggotaan dalam kelompok yang sama berdasarkan tingkat persamaan diantara anggota-anggotanya. Pendekatan lainnya adalah dengan membuat sekumpulan fungsi yang mengukur beberapa properti dari pengelompokan tersebut sebagai fungsi dari beberapa parameter dari sebuah clustering (Witten, 2012). Metode *K-Means* merupakan metode yang termasuk dalam algoritma *clustering* berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada data numerik.

Pengelompokan data dengan metode *K-Means* dilakukan dengan algoritma (Dhuhita, 2015):

1. Tentukan jumlah kelompok
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat kelompok (*centroid*) dari data yang ada di masing-masing kelompok. Lokasi *centroid* setiap kelompok diambil dari rata-rata semua nilai data pada setiap fiturnya. Jika M menyatakan jumlah data sebuah kelompok, i menyarakan fitur ke-i dalam sebuah kelompok, dan p menyatakan dimensi data, maka persamaan untuk menghitung centroid fitur ke-i digunakan persamaan berikut

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j$$

   yang dilakukan sebanyak p dimensi, dari i=1 sampai dengan i=p.
4. Alokasikan masing-masing data ke centroid terdekat. Ada beberapa cara yang dapat dilakukan untuk mengukur jarak data ke pusat kelompok, diantaranya adalah *Euclidean*. Pengukuran jarak pada ruang jarak (*distance space*) *Euclidean* dapat dicari menggunakan persamaan berikut

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

   Data dialokasikan ulang secara tegas ke kelompok yang mempunyai *centroid* dengan jarak terdekat dari data tersebut. Pengalokasian data ini menurutt MacQueen (1967) dapat ditentukan menggunakan persamaan berikut

$$a_i = \begin{cases} 1\,, & d = \min\{D(x_i, c_i)\} \\ 0\,, & lainnya \end{cases}$$

   Fungsi objektif yang digunakan untuk metode *K-Means* ditentukan berdasarkan jarak dan nilai keanggotaan data dalam kelompok. Fungsi objektif menurut McQueen (1967) dapat ditentukan menggunakan persamaan berikut

$$J = \sum_{i=1}^{n}\sum_{i=1}^{k} a_{ic} D(x_i, c_i)^2$$

5. Kembali ke langkah 3, apabila masih terdapat daya yang berpindah kelompok atau apabila terdapat perubahan nilai *centroid* di atas nilai ambang yang ditentukan, atau apabila perubahan nilai pada fungsi objektif yang digunakan masih di atas nilai ambang yang ditentukan.

Pembahasan

Tujuan pada penelitian ini adalah untuk melakukan pengelompokan daerah penghasil padi menurut luas lahan, produksi, dan produktivitas di setiap provinsi di Indonesia menggunakan *K-Means*. Penelitian menggunakan data sekunder berupa data luas lahan, produksi, dan produktivitas padi di setiap provinsi di Indonesia pada tahun 2022 yang bersumber dari Badan Pusat Statistika (BPS). Parameter yang digunakan untuk melakukan pengelompokan daerah penghasil padi di Indonesia berjumlah 3 yaitu luas lahan pertanian padi, produksi padi, dan produktivitas padi.

Jumlah data yang akan digunakan sebanyak 34 data dari setiap provinsi yang ada di Indonesia. Data ini dapat dilihat pada tabel berikut:

Tabel 1. Data Luas Panen, Produksi, dan Produktivitas Tanaman Padi di Indonesia

Provinsi	Luas Lahan (ha)	Produksi (ton)	Produktivitas (kw/ha)
Aceh	271750.20	1509456.00	55.55
Sumatera Utara	411462.10	2088584.00	50.76
Sumatera Barat	271883.10	1373532.00	50.52
Riau	51054.04	213557.20	41.83
Jambi	60539.59	277743.80	45.88
Sumatera Selatan	513378.20	2775069.00	54.06
Bengkulu	57151.84	281610.10	49.27
Lampung	518256.10	2688160.00	51.87
Kep. Bangka Belitung	15107.80	61425.07	40.66
Kep. Riau	179.48	506.91	28.24
DKI Jakarta	477.25	2337.77	48.98
Jawa Barat	1662404.00	9433723.00	56.75
Jawa Tengah	1688670.00	9356445.00	55.41
DI Yogyakarta	110927.20	561699.50	50.64
Jawa Timur	1693211.00	9526516.00	56.26
Banten	337240.70	1788583.00	53.04
Bali	112320.60	680601.60	60.59
Nusa Tenggara Barat	270092.90	1452945.00	53.79
Nusa Tenggara Timur	183092.00	756049.90	41.29
Kalimantan Barat	241478.60	731225.80	30.28
Kalimantan Timur	108226.80	343918.80	31.78
Kalimantan Utara	214908.90	819419.20	38.13
Sulawesi Utara	64970.01	239425.30	36.85
Sulawesi Tengah	8604.19	30533.59	35.49

Sumber: Badan Pusat Statistika

Sulawesi Selatan	58195.56	243730.30	41.88
Sulawesi Tenggara	168993.20	744408.70	44.05
Gorontalo	1038084.00	5360169.00	51.64
Sulawesi Barat	118258.80	478958.00	40.50
Maluku	46823.47	240134.50	51.29
Maluku Utara	69323.95	353513.30	50.99
Papua Barat	23987.82	92601.06	38.60
Papua	6416.45	24486.03	38.16

Sumber: Badan Pusat Statistika

Akan dibentuk 3 *cluster* pada pengelempokan ini, yaitu: (1) *Cluster* 1 Daerah Penghasil Padi Terbesar, (2) *Cluster* 2 Daerah Penghasil padi Menengah, (3) *Cluster* 3 Daerah Penghasil Padi Terendah.

Tabel 2. Hasil *Cluster.*

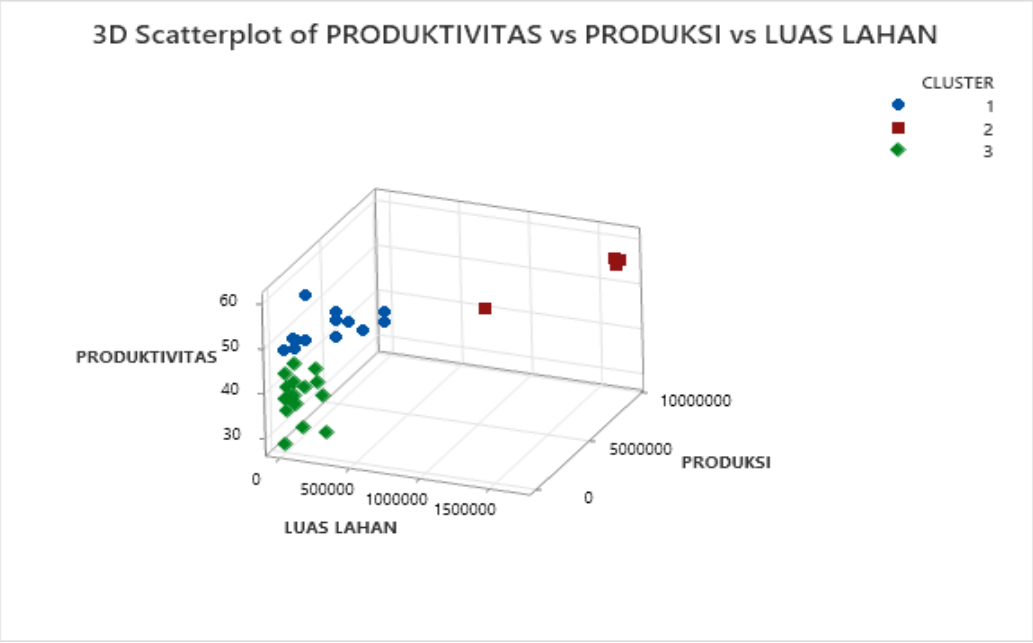
Provinsi	***Cluster***	**Provinsi**	***Cluster***
Aceh	1	Riau	3
Sumatera Utara	1	Jambi	3
Sumatera Barat	1	Sumatera Selatan	1
Bengkulu	1	Kalimantan Barat	3
Lampung	1	Kalimantan Timur	3
Kep. Bangka Belitung	3	Kalimantan Utara	3
Kep. Riau	3	Sulawesi Utara	3
DKI Jakarta	1	Sulawesi Tengah	3
Jawa Barat	2	Sulawesi Selatan	3
Jawa Tengah	2	Sulawesi Tenggara	3
DI Yogyakarta	1	Gorontalo	2
Jawa Timur	2	Sulawesi Barat	3
Banten	1	Maluku	1
Bali	1	Maluku Utara	1
Nusa Tenggara Barat	1	Papua Barat	3
Nusa Tenggara Timur	3	Papua	3

Dari hasil pengujian pada Tabel 4.2 dapat disimpulkan bahwasannya *cluster* 1 terdiri dari 13 provinsi daerah penghasil padi terbesar, *cluster* 2 terdiri dari 4 provinsi daerah penghasil padi menengah, dan *cluster* 3 terdiri dari 17 provinsi daerah penghasil padi terendah ditinjau dari luas lahan, produksi, serta produktivitasnya. Provinsi Aceh dengan luas lahan 271750.20 ha, produksi padi 1509456.00 ton, dan produktivitas padi 55.55 kw/ha masuk ke dalam cluster 1 yaitu kelompok daerah penghasil padi terbesar di Indonesia. Berbeda dengan Provinsi Papua yang memiliki luas lahan 49741.91 ha, produksi padi 193943.50 ton, dan produktivitas padi 38.99 kw/ha masuk ke dalam cluster 3 yaitu kelompok daerah penghasil padi terendah di Indonesia. Adapun jarak antar pusat *cluster* dituangkkan pada tabel berikut.

Tabel 3. Jarak antar pusat cluster

Distance Between Cluster Centroid			
	Cluster 1	**Cluster 2**	**Cluster 3**
Cluster 1	0.0000	3.7974	1.7139
Cluster 2	3.7974	0.0000	4.6732
Cluster 3	1.7139	4.6732	0.0000

Analisis dilakukan dengan mengelompokkan data yang dilakukan oleh algoritma *K-Means* dengan menggunakan visualisasi berupa *3D scatterplot*. Pengelompokan dengan *3D scatterplot* adalah suatu cara untuk mengelompokkan data dengan menggunakan grafik *scatterplot* tiga dimensi. *3D scatterplot* adalah grafik yang menampilkan titik-titik data pada sistem koordinat tiga dimensi, di mana masing-masing sumbu merepresentasikan satu variabel. Dengan menggunakan *3D scatterplot*, kita dapat mengelompokkan data berdasarkan pola atau kecenderungan yang ada pada titik-titik data yang terlihat dalam grafik.



Gambar 3. 3D Scatterplot; x (luas lahan); y (produksi); z (produktivitas)

Dengan melihat pada Gambar 4.2 provinsi-provinsi yang berada pada *cluster* 1 yaitu daerah dengan penghasil padi terbesar cenderung memiliki luas lahan yang sempit namun mereka dapat memanfaatkan lahan tersebut dengan baik sehingga produksi serta produktivitas tanaman padi mereka cukup tinggi. Berbeda dengan *cluster* 2 yaitu daerah dengan penghasil padi menengah memiliki luas lahan yang cukup luas namun produksi serta produktivitas tanaman padi pada *cluster* ini cukup rendah atau setara dengan *cluster* 1. Pada *cluster* 3 yaitu daerah penghasil padi terendah memiliki luas lahan yang sempit, serta produksi dan produktivitas tanaman padi yang juga rendah. Pada *cluster* 3 antara luas lahan berbanding lurus dengan produksi serta produktivitas tanaman padi. Sedangkan pada *cluster* 2 antara luas lahan berbanding terbalik dengan produksi serta produktivitas tanaman padi. Dimana dengan luas lahan yang cukup luas memiliki produksi dan produktivitas yang setara dengan luas lahan yang sempit pada *cluster* 1.

Kesimpulan dan Saran

Kesimpulan dari penelitian ini adalah *K-Means clustering* dapat mengelompokan daerah penghasil padi di Indonesia menurut luas panen, produksi, dan produktivitas padi dengan diperoleh hasil sebesar 13 provinsi tergolong ke dalam daerah penghasil padi terbesar, 4 provinsi yang tergolong ke dalam kelompok daerah penghasil padi menengah, serta 17 provinsi yang tergolong ke dalam kelompok dalam penghasil padi terendah. Pada penelitian ini juga ditemukan bahwa luas lahan tidak menjamin produksi serta produktivitas tanaman padi. Terdapat faktor lain yang perlu dilakukan kajian ulang terkait hal yang mempengaruhi penanaman padi di setiap provinsi. Seperti ditemukan pada penelitian ini, DKI Jakarta yang merupakan kota metropolitan dengan gedung-gedung tinggi dan lahan yang seakan telah penuh mampu tergolong pada daerah penghasil padi terbesar. Meninjau dari luas lahannya yang hanya berkisar 477.25 yang tergolong cukup sempit mampu memiliki produksi serta produktivitas tanaman padi yang tinggi. Dari hasil ini, tentunya ada faktor lain yang dapat meningkatkan produksi dan produktivitas tanaman padi terlepas dari luas lahan yang dimiliki. Oleh karena itu diharapkan ada penelitian lanjutan untuk membahas terkait permasalahan ini.

Daftar Pustaka

Dhuhita, W. M. (2015). Clustering menggunakan metode K-Means untuk menentukan status gizi balita. *Jurnal Informatika*, 160-174.

Metisen, B. M., & Sari, H. L. (2015). Analisis clustering menggunakan metode K-Means dalam pengelompokkan penjualan produk Pasa Swalayan Fadhila. *Jurnal Media Informatika*, 110-118.

Morisette, L., & Chartier, S. (2013). The K-Means clustering technique: General consider and implementation in mathematica. *Tutorial in Quantitive Methods for Psychology*, 15-24.

Muller, A., & Guido, S. (2016). *Introduction to machine learning with Python.* O'Reilly Media.

Rahman, I. F. (2019, July 10). *Metode analisis cluster [Bagian 1]*. Retrieved April 11, 2023, from Medium: https://medium.com/@16611120/metode-analisis-cluster-part-1-eb6c4556d363

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining.* Amerika Serikat: Addison-Wesley Longman Publishing .

Witten. (2012). *Data mining practical machine learning tools and technique 2nd edition.* San Fransisco: Morgan Kaufman Publishers Inc.