Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Pengelompokan Toko Pupuk Termurah *E-commerce* Shopee dengan Metode Klasterisasi

Ni Putu Esti Utami Barsua[1], I Made Joel Jaya Dilaga[2], Rika Lusiana Simbolon[3], Robert Kurniawan[4]
[1, 2, 3]*Program Studi Statistika, Politeknik Statistika STIS, Jakarta, Indonesia*
[4]*Program Studi Komputasi Statistik, Politeknik Statistika STIS, Jakarta, Indonesia*
[1]212112256@stis.ac.id
[2]212112101@stis.ac.id
[3]212112323@stis.ac.id
[4]robertk@stis.ac.id
Corresponding author email: 212112256@stis.ac.id

Abstract: *With a large population and high demand for fertilizers, especially among millennial farmers active on Shopee, it is important to identify high-quality fertilizer stores with affordable prices. This study aims to identify the best method for clustering fertilizer stores on Shopee based on ratings, followers, chat performance, and joining time. The data used is secondary data from 263 fertilizer stores on Shopee, collected on April 27, 2024. The methodology applied is cluster analysis using the K-Means and K-Medoids methods. A Hopkins statistic of 0.934 indicates that the data is suitable for clustering. The Elbow and Silhouette methods were used to identify four optimal clusters. K-Means showed better performance as it met all evaluation criteria, namely the Silhouette coefficient, Dunn index, entropy, Calinski-Harabasz Index (CH Index), and separation index. The clustering results with K-Means recommend the third cluster because it has more followers, high chat performance, and a long joining time. Conversely, the second cluster is not recommended because it has fewer followers, low chat performance, and a short joining time. This study demonstrates that the K-Means method is more effective in clustering fertilizer stores on Shopee.*
Keywords: *K-Means, K-Medoids, fertilizer, cluster*

Abstrak: Dengan jumlah penduduk yang besar dan tingginya permintaan terhadap pupuk, terutama di kalangan petani milenial yang aktif di Shopee, penting untuk mengidentifikasi toko-toko pupuk berkualitas dengan harga terjangkau. Penelitian ini bertujuan mengidentifikasi metode terbaik untuk mengelompokkan toko pupuk di Shopee berdasarkan rating, jumlah pengikut, kinerja *chat*, dan waktu bergabung. Data yang digunakan adalah data sekunder dari 263 toko pupuk di Shopee, diambil pada 27 April 2024. Metodologi yang diterapkan adalah analisis klaster dengan metode *K-Means* dan *K-Medoids*. Statistik *Hopkins* sebesar 0,934 menunjukkan bahwa data cocok untuk klasterisasi. Metode *Elbow* dan *Silhouette* digunakan untuk mengidentifikasi empat klaster optimal. *K-Means* menunjukkan performa lebih baik karena memenuhi semua kriteria evaluasi, yaitu *Silhouette coefficient*, *Dunn index*, *entropy*, *Calinski-Harabasz Index* (*CH Index*), dan *separation index*. Hasil klasterisasi dengan *K-Means* menunjukkan bahwa klaster ketiga direkomendasikan karena memiliki lebih banyak pengikut, performa *chat* yang tinggi, dan waktu bergabung yang lama. Sebaliknya, klaster kedua tidak direkomendasikan karena memiliki pengikut yang sedikit, performa *chat* yang rendah, dan waktu bergabung yang singkat. Penelitian ini menunjukkan bahwa metode *K-Means* lebih efektif dalam mengelompokkan toko pupuk di Shopee.
Kata kunci: *K-Means*, *K-Medoids*, pupuk, *cluster*

I. PENDAHULUAN

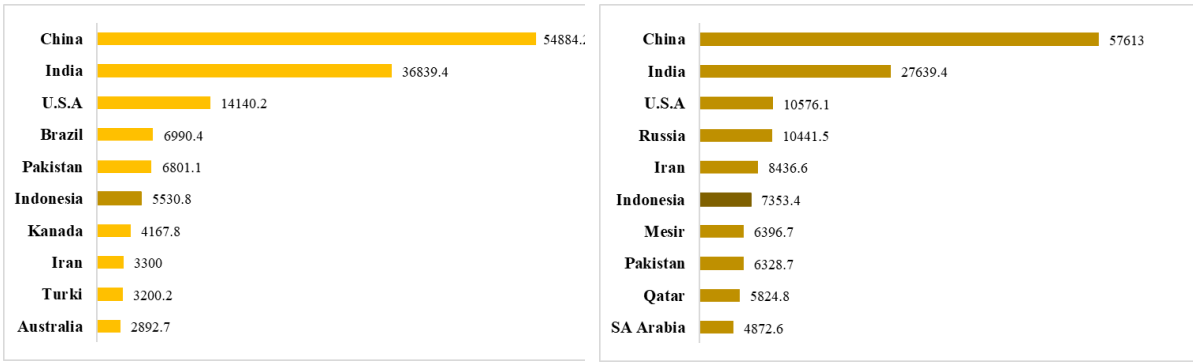
Saat ini, teknologi berkembang dengan pesat dan berhasil mengubah gaya hidup penduduk dunia, termasuk dalam hal belanja yang kini telah beralih ke belanja *online*. Pandemi Covid-19 juga berkontribusi mempercepat perubahan perilaku belanja melalui toko *online*. [1]. Terdapat beberapa platform belanja *online* di Indonesia, salah satunya Shopee. Pada tahun 2023, Shopee menduduki peringkat pertama dengan total kunjungan mencapai 2,35 miliar. [2]. Jumlah tersebut melampaui pesaingnya seperti Tokopedia, yang menerima 1,25 miliar kunjungan, dan Lazada dengan 762,4 juta kunjungan.

Indonesia termasuk penduduk dengan populasi terbesar, yakni posisi keempat dunia. Generasi muda, terutama generasi milenial dan Gen Z, mendominasi demografi penduduk Indonesia. Berdasarkan hasil Sensus Penduduk persentase generasi Z dan milenial di Indonesia adalah 52,81%

atau setara dengan 145,39 juta jiwa dari total populasi berjumlah 270,2 juta jiwa [3]. Platform belanja daring asal Singapura, Shopee, menjadi pilihan utama 69,9% responden dari kalangan generasi Z dan 64,2% responden generasi milenial, menjadikannya platform *e-commerce* Shoppe yang paling banyak dikunjungi oleh kelompok usia ini [4].

Selain itu, Indonesia juga memiliki julukan sebagai Negara Agraris, menandakan pentingnya sektor pertanian bagi perekonomian nasional. Berdasarkan Renstra Kementerian Pertanian tahun 2020-2024, pertanian di Indonesia berperan vital dalam menyediakan lapangan kerja dan memenuhi kebutuhan pangan [5]. Oleh karena itu, sektor ini memerlukan berbagai produk pendukung, termasuk yang paling krusial seperti pupuk. Pupuk menjadi elemen kunci dalam meningkatkan hasil panen dan kualitas produk pertanian, yang pada gilirannya berdampak positif pada ketahanan pangan dan stabilitas ekonomi negara [6].

Berdasarkan Peraturan Menteri Pertanian Nomor 4 Tahun 2019 tentang Pedoman Gerakan Pembangunan Sumber Daya Manusia Pertanian Menuju Lumbung Pangan Dunia 2045, petani milenial adalah petani berusia 19 tahun sampai 39 tahun, dan/atau petani yang adaptif terhadap teknologi digital [7]. Peningkatan jumlah petani milenial juga menunjukkan tren positif di sektor ini. Pada tahun 2023, Badan Pusat Statistik mencatat terdapat 6,18 juta petani millennial dan 2,6 juta diantaranya menggunakan teknologi digital seperti alsintan modern, internet, telepon pintar, dan *drone* [8]. Tidak hanya petani muda yang adaptif, sebanyak 10,59 juta petani dengan usia lebih dari 39 tahun menggunakan teknologi digital dalam pekerjaannya [8].



Gambar 1a. **Gambar 1b.**

Gambar 1. Statistik Pupuk dari 10 Negara Dunia Tahun 2023 (juta ton) : (1a) Konsumsi Pupuk (1b) Produksi Pupuk

Gambaran situasi yang disajikan pada Gambar 1 menggambarkan peran Indonesia yang signifikan dalam produksi dan konsumsi pupuk di dunia pada tahun 2023. Data menunjukkan bahwa Indonesia memperoleh peringkat 6 besar dalam hal produksi dan konsumsi pupuk, dengan produksi sebesar 7,3 juta ton dan konsumsi mencapai 5,5 juta ton pada tahun tersebut. Pupuk berperan penting dalam meningkatkan produktivitas tanaman, menyumbang sekitar 20-40% peningkatan produksi. Lebih jauh, pupuk dan sistem irigasi merupakan faktor utama yang berpengaruh terhadap hasil produksi padi, salah satu komoditas pertanian kunci di Indonesia. Dengan demikian, penggunaan dan ketersediaan pupuk memiliki implikasi yang signifikan terhadap pertumbuhan sektor pertanian dan keamanan pangan di Indonesia [9].

Dengan jumlah penduduk yang besar dan tingginya permintaan terhadap pupuk, terutama di kalangan petani milenial yang aktif di platform *e-commerce* Shopee, penting untuk mengidentifikasi toko-toko pupuk yang menawarkan produk berkualitas dengan harga terjangkau. Hal ini dikarenakan pada 2013-2021 terdapat tren positif pada jumlah toko online pada *marketplace* yang menjual khusus

produk pertanian, hingga mencapai 300 toko yang tersebar pada berbagai platform *e-commerce* [10]. Oleh karena itu, diperlukan suatu sistem pengelompokan yang dapat memudahkan konsumen dalam menilai mutu masing-masing toko yang menjual pupuk di Shopee.

Clustering adalah teknik untuk mengelompokkan amatan ke dalam beberapa klaster berdasarkan atribut tertentu. Prinsip dasar dari teknik ini adalah untuk memastikan bahwa objek dalam grup yang sama memiliki banyak kesamaan, sementara objek di grup yang berbeda memiliki perbedaan yang signifikan. Hasil clustering yang baik dicapai ketika kesamaan antar objek dalam satu grup tinggi dan kesamaan antar grup rendah [17]. Dua metode *clustering* yang sering digunakan untuk data tanpa label adalah *K-Means* dan *K-Medoids*. Kedua metode ini bertujuan untuk membagi kumpulan data menjadi *k* klaster. Namun, perbedaannya terletak pada cara menentukan pusat klister. *K-Means* menggunakan nilai rata-rata data dalam setiap klaster untuk menetapkan *centroid*, sedangkan *K-Medoids* memilih titik pusat dari data yang ada secara acak [18].

Berdasarkan penelitian sebelumnya, disimpulkan bahwa pengelompokan dengan pendekatan *K-Medoids* merupakan metode yang paling unggul [11]. Penelitian serupa dengan amatan yang diklasterkan berupa toko kaus termurah juga menghasilkan kesimpulan yang sama, yakni metode *K-Means* yang lebih baik [12]. Namun, penelitian lain yang menggunakan metode *Clustering* menyatakan bahwa *K-Means* adalah metode terbaik untuk pengelompokan data [13]. Oleh karena itu, dalam kasus ini perlu dilakukan perbandingan untuk menentukan metode yang lebih efektif dalam mengelompokkan toko pupuk berdasarkan beberapa karakteristik. Selain itu, pengelompokan toko produk pupuk termurah di Shopee belum pernah dieksplorasi oleh peneliti sebelumnya.

Penelitian ini bertujuan untuk mengidentifikasi metode yang paling sesuai untuk mengelompokkan toko pupuk di Shopee berdasarkan rating, jumlah pengikut, kinerja *chat*, dan waktu bergabung. Dengan pengelompokan ini, toko-toko pupuk akan terorganisir dalam kelompok yang bervariasi reputasinya berdasarkan variabel yang ditentukan. Hal ini dapat berfungsi sebagai sumber informasi bagi pelanggan untuk memilih toko pupuk berkualitas dengan harga terjangkau serta membangun kepercayaan terhadap penjual. Harapannya studi ini dapat membantu pembeli menemukan toko yang sesuai dengan preferensi dan kebutuhan mereka, sehingga memberikan manfaat yang signifikan bagi konsumen pupuk.

II. METODE PENELITIAN

Penelitian ini menerapkan metode klasterisasi dengan prosedur rinci dijelaskan dalam sub bab berikut.

2.1. Data dan Sumber Data

Penelitian ini memanfaatkan data sekunder berupa *dataset* karakteristik pada 263 toko pupuk. Data yang digunakan bersumber dari *e-commerce* Shopee yang berhasil diambil pada tanggal 27 April 2024. Karakteristik toko yang berhasil dikumpulkan mencakup banyak pengikut, rating toko, performa *chat*, dan waktu bergabung.

2.2. Preprocessing Data

Terdapat 944 baris data produk dari 278 toko pupuk yang berhasil dikumpulkan. Namun, data ini masih mengandung *missing value* dan inkonsistensi. Hal ini dapat menurunkan akurasi hasil yang akan diperoleh sehingga perlu dilakukan *preprocessing*. Pada tahap ini dilakukan pengecekan *missing value*, *outlier*, standardisasi, serta pengecekan multikolinieritas pada data. Untuk *missing value* dilakukan penghapusan baris data sehingga data yang akan digunakan menjadi 429 baris produk dari 263 toko pupuk. Selain itu juga dilakukan penghapus kolom data yang tidak dibutuhkan pada penelitian.

Berdasarkan hasil pengecekan *outlier* menggunakan *boxplot* pada Rstudio, *dataset* tidak mengandung *outlier*. Selain itu juga terdapat perbedaan satuan pada variabel waktu bergabung, sehingga seluruh data disamakan satuannya menjadi tahun. Pada *dataset* juga terdapat perbedaan satuan antar variabel sehingga perlu dilakukan standardisasi data. Standardisasi data akan membuat semua variabel berada pada skala yang sama sehingga perhitungannya menjadi lebih valid. Standardisasi dilakukan mengikuti Persamaan (1), (2), dan (3).

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{1}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}\right)^2 \tag{2}$$

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_i} \tag{3}$$

Dimana,

N	:	banyak observasi
x_{ij}	:	data ke-i pada variabel ke-j dengan $i = 1,2,\dots,N$ dan $j = 1,2,\dots,n$
$\bar{x}_{ij}$	:	rata-rata pada variabel ke-j
$\hat{x}_{ij}$	:	standardisasi data ke-i variabel ke-j
S_j^2	:	varians variabel ke-j
S_j	:	standar deviasi variabel ke-j

2.3. Statistik Hopkins

Statistik *Hopkins* digunakan untuk menentukan apakah *dataset* yang digunakan memiliki struktur yang cocok untuk dianalisis dengan metode klasterisasi [14]. Statistik *Hopkins* dihitung menggunakan Persamaan (4) dimana H merupakan nilai statistik *Hopkins*, y_i merupakan jarak tetangga terdekat dari *dataset* acak, x_i merupakan jarak tetangga terdekat dari *dataset* asli, dan n adalah jumlah titik sampel dari *dataset.* [14].

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{4}$$

Hipotesis nol menunjukkan bahwa *dataset* asli yang berdistribusi secara seragam (artinya, tidak ada klaster yang bermakna). Hipotesis alternatif menunjukkan bahwa *dataset* tidak didistribusikan secara seragam atau terdapat klaster yang bermakna. Jika nilai statistik *Hopkins* mendekati 1, maka hipotesis nol dapat ditolak dan disimpulkan bahwa terdapat kemampuan klaster (*clusterability)* yang signifikan.

2.4. Penentuan Jumlah Klaster Optimum

Pada penelitian ini, penentuan jumlah *cluster* menggunakan metode *elbow* dan *silhouette*. *Elbow Method* merupakan salah satu metode tertua untuk menentukan jumlah *cluster* yang optimal pada *dataset* yang dianalisis [15]. Dasar metode ini adalah dengan mengatur $k = 2$ sebagai jumlah *cluster* optimal awal, kemudian terus menaikkan nilai k hingga mencapai nilai maksimal yang ditetapkan untuk jumlah *cluster* optimal potensial yang diestimasi. *Silhouette Method* adalah metode lain yang terkenal

dengan kinerja yang baik dalam memperkirakan jumlah *cluster* optimal potensial. Metode ini memanfaatkan rata-rata jarak antara satu titik data dengan titik lain dalam *cluster* yang sama, serta rata-rata jarak antara *cluster* yang berbeda untuk menilai hasil pengelompokan.

2.5. Clustering

Clustering merupakan upaya mengumpulkan data serupa menjadi satu grup dan data tersebut tidak diperbolehkan muncul di grup lain [16]. Biasanya, *clustering* dapat dibedakan menjadi dua jenis utama, yaitu pengelompokan hierarki dan pengelompokan partisi. Pengelompokan hirarki membagi pola secara bertahap menggunakan pendekatan dari bawah ke atas (*agglomerative*) atau dari atas ke bawah (*divisive*). Dalam pengelompokan hierarki, klaster terbentuk dengan membagi pola secara berulang menggunakan pendekatan dari atas ke bawah atau dari bawah ke atas. Sebaliknya, pengelompokan partisi berusaha mengklasifikasikan pengamatan data ke dalam k klaster berdasarkan beberapa fungsi kriteria. Fungsi kriteria yang umum dicoba adalah menemukan jarak minimum antara titik dalam klaster yang tersedia. Algoritma seperti *K-Means*, *K-Medoids*, dan CLARA termasuk dalam kategori ini. CLARA dirancang untuk menangani data dengan ribuan objek, sehingga tidak digunakan dalam eksperimen saat ini. Namun, untuk menginvestigasi efektivitas metode pengelompokan serupa pada kumpulan data besar, penelitian masa depan tetap diperlukan. Meskipun demikian, set data yang digunakan dalam penelitian ini berukuran kecil hingga sedang, sehingga aplikasi praktis yang dikembangkan dengan metodologi yang diusulkan akan mengurangi waktu komputasi saat menghasilkan hasil. Diskusi lebih lanjut pada topik ini difokuskan pada *K-Means* dan *K-Medoids*.

2.6. K-Means

K-Means adalah algoritma sederhana yang dikenal karena keefektifannya dalam melakukan pengelompokan [16]. Tujuan utama dari *K-Means* adalah untuk mengurangi perbedaan kesalahan *error* kuadrat antara pusat klaster dan titik data di dalam klaster tersebut. Untuk beberapa titik data n dimensi yang ingin dikelompokkan dalam k klaster dengan μ_k sebagai rata-rata dari klaster tersebut, maka *K-Means* dapat dijelaskan dengan Persamaan (5).

$$\left[\sum_{k=1}^{k}\sum_{x\in c_k}\left||x_i-\mu_k|\right|^2\right] \tag{5}$$

Dimana x_i merupakan set data *point* dengan $i = 1,2,3,\ldots,n$ yang akan dikelompokkan ke dalam klaster ck dengan $k = 1,2,3,..,k$. Untuk mengurangi kesalahan kuadrat, *K-Means* menempatkan pola ke dalam klaster k yang awalnya terbagi secara partisi. Langkah-langkah membentuk klaster dengan metode *K-Means* [17]. Langkah-langkah membentuk klaster dengan metode *K-Means* adalah sebagai berikut [17] :

1. Menentukan jumlah klaster (k) yang akan dibentuk.
2. Menginisiasi *centroid* secara random.
3. Menghitung jarak tiap observasi terhadap tiap *centroid* dengan metode *euclidean distance.* Rumus *euclidean distance* dapat dituliskan dengan Persamaan (6).

$$d_{(x,y)} = |x-y| = \sqrt{\left(\sum_{i=1}^{n}(x_i-y_i)\right)^2} \tag{6}$$

Dimana,

$d_{(x,y)}$: jarak *euclidean* antara observasi terhadap *centroid*

x_i : data observasi ke-i
y_i : *centroid* pada *cluster* ke-i

4. Hitung nilai rata-rata masing-masing klaster dan gunakan sebagai *centroid* yang baru.
5. Hitung kembali jarak tiap observasi terhadap *centroid* yang baru dan masukkan observasi ke dalam klaster yang memiliki jarak terdekat.
6. Iterasi akan berhenti ketika hasil yang didapatkan sama dengan iterasi sebelumnya.

2.7. K-Medoids

K-Medoids atau partisi sekitar algoritma *medoids* adalah variasi dari algoritma *K-Means* [16]. Berbeda dengan *K-Means* yang memilih mean sebagai *centroid*, dalam *K-Medoids* titik data dipilih sebagai *medoid*. *Medoid* ini dapat dianggap sebagai objek dalam klaster yang memiliki ketidaksamaan rata-rata minimal dengan objek lain dalam klaster tersebut. Proses algoritma *K-Medoids* dimulai dengan menghitung k *medoid* dan menugaskan setiap objek data ke *medoid* terdekat menggunakan beberapa metrik jarak. Setelah itu, *K-Medoids* menghitung biaya (*cost)* pertukaran untuk objek pertukaran P_i dan medoid M_i dengan rumus dalam Persamaan (7) berikut.

$$Cost_{pm} = \sum_{M_i} \sum_{p_i \in m_i} |P_i - M_i| \tag{7}$$

Untuk setiap medoid M, titik data P sedemikian rupa sehingga $P \neq M$,

1. Dalam pertukaran antara M dan P, hitung perubahan nilai *cost.*
2. Jika perubahan *cost* adalah semakin menurun, maka nilai M ditukar dengan P.
3. Lakukan pertukaran M dan P. Jika, hal ini mengurangi *cost* lalu ulangi langkah 1 dan 2. Jika tidak, algoritma akan berhenti

2.8. Metode Evaluasi Klaster

Setelah dilakukan *Clustering,* hasil yang terbentuk akan dievaluasi dengan metode berikut.

1. Silhouette Index

Banyak metode evaluasi kinerja *cluster* yang memerlukan set pelatihan, tetapi indeks *Silhouette* tidak memerlukan hal tersebut untuk mengevaluasi hasil pengelompokan [18]. Nilai indeks *Silhouette* berada di antara -1 dan 1. Jika nilainya negatif berarti $a(x_i)$ lebih besar dibandingkan $b(x_i)$ dan *within dissimilarity* lebih besar dibandingkan *between dissimilarity*. Lebar *Silhouette* $s(x_i)$ untuk titik data x_i didefinisikan sebagai Persamaan (8) dan (9).

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i), a(x_i)\}} \tag{8}$$

$$b(x_i) = \min\{d_i(x_i)\} \tag{9}$$

Dimana,

x_i : elemen dalam *cluster* k
$a(x_i)$: rata-rata jarak x_i terhadap elemen lain dalam *cluster* k (*within dissimilarity)*
$b(x_i)$: nilai minimum $d_i(x_i)$ diantara seluruh klaster $l \neq k$
$d_i(x_i)$: rata-rata jarak x_i terhadap seluruh titik di dalam *cluster* l untuk $l \neq k$ (*between dissimilarity)*

2. Dunn Index

Indeks *Dunn* (DU) didefinisikan sebagai rasio antara jarak minimum antar objek data dari *cluster* yang berbeda dan yang terbesar di dalamnya jarak cluster [18]. Indeks *Dunn* memiliki nilai antara 0 sampai ∞ dan tujuannya adalah untuk menghasilkan nilai Indeks *Dunn* yang paling besar. Indeks *Dunn* dijelaskan sebagai Persamaan (10).

$$DU = \frac{d_{min}}{\mathrm{d_{max}}} \tag{10}$$

Dimana,

d_{min} : jarak minimum antara dua objek data dari *cluster* yang berbeda
d_{max} : jarak maksimum antara dua objek data dari *cluster* yang sama

3. *Classification Entropy*

Indeks ini dapat didefinisikan menggunakan fungsi matematis pada Persamaan (11) [7].

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{i,k}\ln\left(\mu_{i,k}\right) \tag{11}$$

Dimana,

$\mu_{i,k}$: fungsi keanggotaan i
$\mathrm{d_{max}}$: jumlah titik data dan jumlah *cluster*

4. *Calinski-Harabasz Index* (*CH Index*)

Indeks Calinski-Harabasz (*CH Index*) adalah salah satu metode evaluasi yang digunakan untuk mengukur kualitas pengelompokan dalam analisis klaster. Tujuannya adalah untuk mengevaluasi seberapa baik sebuah pengelompokan memisahkan antara klaster-klaster yang berbeda dan seberapa padat *(compact)* setiap klaster tersebut [19]. Untuk n titik data dan k *cluster* dapat dituliskan sebagai Persamaan (12) dan (13).

$$\frac{\left[\frac{trace\ (B)}{K-1}\right]}{\left[\frac{trace(W)}{n-K}\right]} \tag{12}$$

$$trace\ W = \sum_{k=1}^{K}\sum_{i=1}^{n_k} n_k\left|\left|x_i - z_k\right|\right|^2 \tag{13}$$

Dimana,

B : *between scatter matrices*
W : *within scatter matrices*

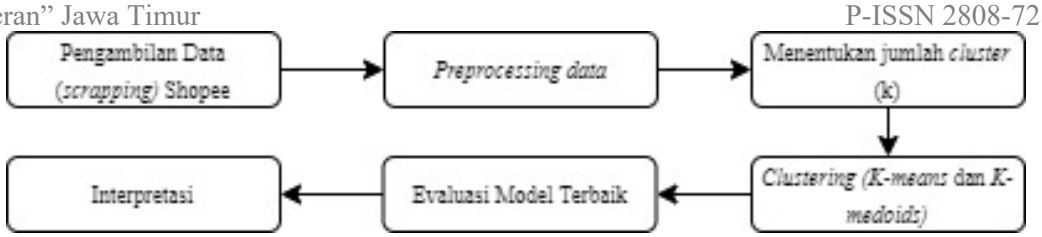
5. *Separation Index*

Indeks ini dapat didefinisikan menggunakan fungsi matematis pada Persamaan (14) [7]. *Centroid* dari *cluster* i dan j direpresentasikan sebagai v_i dan v_j. N dan c adalah jumlah titik data dan jumlah *cluster.*

$$S(c) = \frac{1}{N}\sum_{i=1}^{c}\frac{\sum_{k=1}^{N}\mu_{i,k}^2\left|\left|x_k - v_i\right|\right|^2}{\min\limits_{i,j,i\neq j}\left|\left|v_j - v_i\right|\right|^2} \tag{14}$$

2.9. *Tahapan Penelitian*

Langkah-langkah penelitian dari awal sampai akhir tergambar pada diagram alir pada Gambar 2.



Gambar 2. Diagram Alir Penelitian

III. HASIL DAN PEMBAHASAN

Dataset yang terkumpul merupakan *dataset* hasil *scraping* sehingga membutuhkan tahapan-tahapan pengolahan. Pengolahan data menggunakan *software* Excel dan RStudio. Di bawah ini disajikan proses dan hasil serta pembahasan yang telah dilakukan.

3.1. Standardisasi Data

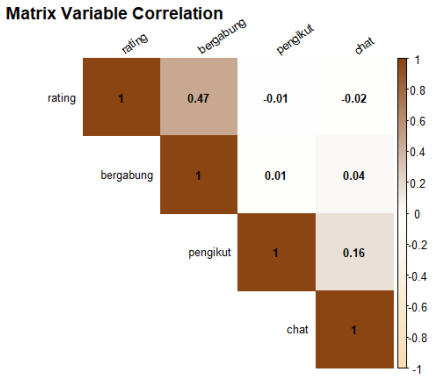
Dataset yang digunakan merupakan data toko pupuk termurah di Shopee yang terdiri dari empat variabel. Dari keempat variabel, terdapat variabel yang memiliki satuan berbeda, sehingga perlu dilakukan standardisasi data menggunakan *z-score* untuk menyamakan satuan. Variabel-variabel sebelum distandarisasi disajikan dalam Tabel 1.

Tabel 1. Keterangan Variabel

Nama variabel	Keterangan	Satuan
rating	Rating toko	-
bergabung	Waktu bergabung	Tahun
pengikut	Jumlah pengikut toko	Akun
chat	Performa *chat* dibalas	Persen

3.2. Pemeriksaan Non Multikolinearitas

Pengecekan asumsi ini menggunakan matriks korelasi antar variabel yang disajikan dalam Gambar 3. Berdasarkan hasil yang diperoleh, tidak terdapat indikasi adanya hubungan antara dua variabel yang disajikan dalam bentuk korelasi pearson. Nilai korelasi tersebut tidak ada yang melebihi 0.8 yang mengindikasikan bahwa asumsi non multikolinearitas terpenuhi [20]. Oleh karena itu, penghilangan variabel tidak diperlukan.



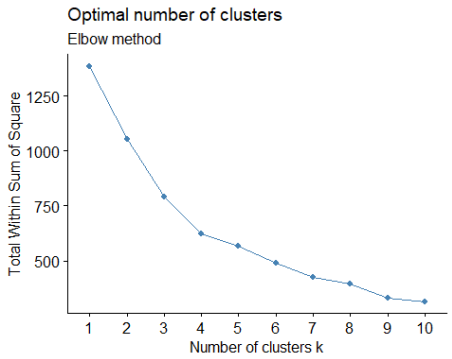
Gambar 3. Pemeriksaan Asumsi Non Multikolinearitas

3.3. Statistik Hopkins

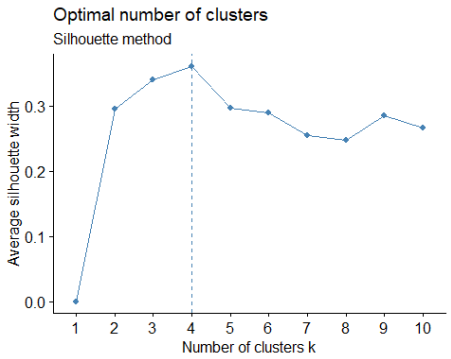
Langkah awal sebelum melakukan analisis klaster adalah memeriksa *Hopkins Statistics*. Dengan menggunakan *software* R diperoleh nilai Statistik *Hopkins* sebesar 0,934. Nilai ini bernilai lebih dari 0,5 atau mendekati 1 yang bermakna terdapat kemampuan klaster (*clusterability*) yang signifikan. Oleh karena itu, data cenderung terklaster sehingga dapat dilanjutkan dengan analisis klaster.

3.4. Jumlah Klaster Optimum

Sebelum dilakukan klasterisasi toko pupuk, perlu ditentukan jumlah k optimum dengan menggunakan metode *Elbow* dan metode *Silhouette*. Dalam metode *Elbow*, *Within-cluster Sum of Squares* (WCSS) dihitung untuk berbagai jumlah klaster k. lalu, nilai WCSS ini dipetakan dalam sebuah grafik terhadap jumlah klaster. Grafik ini menunjukkan penurunan WCSS yang tajam pada awalnya dan kemudian melandai. Titik dimana penurunan ini mulai melambat secara signifikan dan membentuk "siku" menandakan jumlah klaster yang optimum. Dari metode tersebut diperoleh hasil seperti pada Gambar 4a. Penggunaan metode *Elbow* memiliki kelemahan yakni subjektivitas interpretasi k. Oleh karena itu, tetap dilakukan metode *Silhouette.* Metode ini mengukur kualitas dan kecocokan seberapa baik setiap sampel data cocok dalam klasternya. Dari metode tersebut diperoleh hasil seperti pada Gambar 4b.



Gambar 4a.



Gambar 4b.

Gambar 4. Jumlah klister optimum dari metode : (4a) *Elbow* (4b) *Silhouette*

Berdasarkan grafik pada Gambar 4a, jumlah klaster optimum ini ditandai dengan titik *elbow* yang terlihat jelas menunjukkan penambahan jumlah klaster setelah titik tersebut tidak memberikan pengurangan WCSS yang signifikan lagi. Dalam hal ini, titik *elbow* tercapai ketika $k = 4$. Sedangkan, pada Gambar 4b menunjukkan jumlah klaster optimum ditandai dengan nilai *average silhouette* yang paling tinggi atau memiliki tingkat kohesi yang baik di dalam klaster dan tingkat pemisahan yang baik antara klaster. Dalam hal ini, nilai *Average Silhouette* paling tinggi dicapai ketika $k = 4$, sama seperti pada metode *Elbow*. Maka, klasterisasi data menjadi empat klaster memberikan hasil yang lebih baik dibandingkan dengan jumlah klaster lainnya.

3.5. Analisis Klaster dengan K-Means dan K-Medoids

Dengan menggunakan metode *K-Means* dan *K-Medoids*, data akan dikelompokkan sebanyak empat klaster. Gambar 5 menampilkan hasil visualisasi dari kedua metode klasterisasi. Hasil ini memberikan gambaran yang jelas tentang distribusi toko pupuk dalam setiap klaster.

E-ISSN 2808-5841
P-ISSN 2808-7283

Gambar 5a. **Gambar 5b.**

Gambar 5. Visualisasi *cluster plot* dengan metode : (4a) *K-Means* (4b) *K-Medoids*

Berdasarkan Gambar 5 terlihat bahwa terdapat perbedaan mencolok antara metode *K-Means* dan *K-Medoids* dalam hal distribusi observasi. Metode *K-Means* menunjukkan bahwa klaster ketiga memiliki jumlah observasi terbesar, yakni 97 observasi. Klaster keempat memiliki 70 observasi, klaster kedua 52 observasi, dan klaster pertama 44 observasi. Sebaliknya, metode *K-Medoids* menunjukkan bahwa klaster keempat memiliki jumlah observasi terbanyak dengan 86 observasi, diikuti oleh klaster kedua dengan 69 observasi, klaster ketiga dengan 63 observasi, dan klaster pertama dengan 45 observasi. Perbedaan distribusi observasi ini disebabkan oleh perbedaan cara penentuan titik *centroid* pada masing-masing metode.

3.6. Evaluasi Model Terbaik

Klaster dianggap baik dalam mengelompokkan data ketika variasi dalam klaster rendah, tetapi variasi antara klaster tinggi. Untuk menentukan metode terbaik di antara kedua metode yang digunakan sebelumnya, perlu dilakukan evaluasi model. Dalam penelitian ini, ukuran yang digunakan untuk evaluasi antara lain *silhouette coefficient, dunn index, entropy, Calinski-Harabasz Index* (*CH Index*), dan *separation index*. Dengan menggunakan sejumlah metrik evaluasi ini, maka dapat dibandingkan dua metode klasterisasi yaitu *K-Means* dan *K-Medoids* sehingga dapat dipilih metode yang paling optimal dalam memisahkan dan mengelompokkan data dengan presisi yang tinggi. Ukuran evaluasi dari metode *K-Means* dan *K-Medoids* disajikan pada Tabel 2.

Tabel 2. Ukuran Evaluasi *K-Means* dan *K-Medoids*

Ukuran evaluasi	*K-Means*	*K-Medoids*
Silhoutte	0.253	0.225
Dunn	0.020	0.019
Entropy	1.353	1.358
CH Index	81.830	80.008
Separation Index	0.403	0.359

Berdasarkan hasil yang terdapat dalam Tabel 2, disampaikan komparasi antara metode *K-Means* dan *K-Medoids* dengan menggunakan beberapa ukuran evaluasi. Dari metode evaluasi tersebut, terlihat bahwa nilai *entropy K-Means* lebih rendah daripada *K-Medoids*. Sementara itu, jika dilihat dari nilai evaluasi lainnya, seperti *silhouette coefficient, dunn index, Calinski-Harabasz Index* (*CH Index*), dan *separation index*, nilai *K-Means* lebih tinggi dibandingkan *K-Medoids*. Jadi, secara umum dapat disimpulkan, metode *K-Means* menunjukkan performa yang lebih baik daripada *K-Medoids* karena memenuhi semua kriteria dalam ukuran evaluasi yang digunakan. Hal ini mengindikasikan bahwa *K-Means* lebih efektif dalam melakukan klasterisasi data yang ada.

3.7. Analisis Klaster berdasarkan Metode Terbaik

Dari hasil penilaian, metode optimal adalah *K-Means* dengan jumlah klaster optimal *k* adalah empat. Metode ini menunjukkan performa terbaik dibandingkan metode lainnya, yakni *K-Medoids*. Deskripsi masing-masing klaster dapat ditemukan dalam Tabel 3.

Tabel 3. Output *K-Means Clustering*

Klaster	Rating	Bergabung (Tahun)	Pengikut	Performa *Chat* (%)
1	4,4235	2,2398	5984	78,4091
2	4,8859	4,4264	5748	49,9423
3	4,8475	4,2054	25716	88,6186
4	4,8550	6,6114	6853	84,6571

Berdasarkan Tabel 3 di atas, klaster ketiga memiliki rata-rata jumlah pengikut terbanyak dan performa *chat* tertinggi dibandingkan dengan klaster pertama, kedua, dan keempat. Namun, klaster kedua memiliki rata-rata rating tertinggi di antara keempat klaster. Di sisi lain, klaster keempat memiliki rata-rata waktu bergabung yang paling lama dibandingkan dengan klaster lainnya. Klaster pertama, meskipun memiliki jumlah pengikut dan performa *chat* yang lebih rendah, tetap menunjukkan pola interaksi yang menarik untuk diperhatikan. Setiap klaster menunjukkan karakteristik yang berbeda dalam hal jumlah pengikut, performa *chat*, rating, dan waktu bergabung, yang memberikan wawasan beragam tentang dinamika antara penjual dan pembeli di klaster toko pupuk ini.

Tabel 3 menunjukkan bahwa klaster yang paling memenuhi kriteria adalah klaster ketiga. Klaster ini memiliki jumlah pengikut terbanyak (25716), performa *chat* terbaik (88,6186%), dan rating yang sangat tinggi (4,8475), meskipun bukan yang tertinggi. Rating tertinggi dimiliki oleh klaster kedua, tetapi nilainya tidak jauh berbeda dengan klaster ketiga, yakni sama-sama di atas 4,8. Ini menunjukkan bahwa toko-toko dalam klaster ketiga tidak hanya populer tetapi juga sangat responsif dan mendapatkan penilaian tinggi dari pelanggan. Oleh karena itu, klaster ketiga lebih direkomendasikan untuk membeli pupuk jika didasarkan pada harga termurah.

Dilihat dari jumlah pengikut, hal ini mencerminkan tingkat kepercayaan pelanggan terhadap toko. Semakin banyak pengikut yang dimiliki, semakin tinggi pula tingkat kepercayaan yang diberikan kepada toko tersebut [21]. Waktu bergabung yang lebih lama menunjukkan bahwa toko tersebut lebih berpengalaman. Gomez et al. (2022) melakukan studi yang menyatakan bahwa waktu masuk pasar merupakan faktor penting yang memengaruhi kinerja. Semakin lama sebuah perusahaan berada di pasar, semakin sedikit pelanggan yang mengikutinya, dan kinerjanya cenderung menurun karena harus bersaing dengan para pelopor [22].

Selain itu, performa *chat* yang lebih tinggi menandakan layanan yang lebih berkualitas. Studi terdahulu mengungkapkan bahwa *chat* berdampak besar pada keputusan pembelian konsumen. Penelitian ini menemukan bahwa konsumen yang berinteraksi melalui *chat* lebih cenderung untuk melakukan pembelian. Peluang ini meningkat ketika penjual merespons pelanggan dengan cepat. [23].

Menurut Tabel 3, klaster yang paling tidak direkomendasikan ialah klaster kedua. Meskipun klaster kedua memiliki rating tertinggi (4,8859), tetapi klaster ini mempunyai performa *chat* terendah (49,9423%) dan jumlah pengikut yang paling sedikit (5748). Performa *chat* yang rendah bisa menjadi indikator bahwa toko-toko dalam klaster ini kurang responsif terhadap pertanyaan atau masalah pelanggan, yang bisa menyebabkan ketidakpuasan.

Hal ini dapat mengindikasikan bahwa klaster kedua terdiri dari toko-toko yang relatif baru dalam industri ini, mungkin masih dalam tahap awal pengembangan, atau belum berhasil membangun basis pengikut yang signifikan. Kemungkinan besar, toko-toko dalam klaster ini belum memiliki reputasi yang mapan atau belum memperoleh pengalaman yang cukup dalam memenuhi kebutuhan pelanggan.

Dengan waktu bergabung yang relatif singkat dan jumlah pengikut yang masih rendah, klaster kedua mungkin masih dalam proses membangun kehadiran mereka di pasar pupuk.

Berdasarkan jumlah observasi pada klaster pertama, terdapat jumlah yang terbatas, hanya sebanyak 44 observasi. Hal ini menunjukkan bahwa hanya sedikit toko yang menawarkan pupuk berkualitas tinggi dengan harga murah. Temuan ini sejalan dengan penelitian Guizzardi et al. (2022), yang menyimpulkan adanya hubungan positif antara kualitas dan harga. Artinya, produk dengan harga yang lebih tinggi biasanya menawarkan kualitas yang lebih baik. Sebaliknya, produk dengan harga lebih rendah cenderung jarang memiliki kualitas yang unggul [24].

IV. KESIMPULAN

Metode *K-Means* dengan empat klaster terbukti terbaik untuk mengelompokkan 263 toko pupuk termurah di Shopee, dengan klaster ketiga lebih direkomendasikan karena memiliki pengikut lebih banyak, performa *chat* lebih tinggi, dan waktu bergabung lebih lama. Sebaliknya, klaster kedua kurang direkomendasikan karena pengikut lebih sedikit, performa *chat* dan rating lebih rendah, serta waktu bergabung lebih singkat. Klaster pertama memiliki sedikit observasi karena hubungan positif antara harga dan kualitas. Penelitian selanjutnya disarankan menambahkan variabel terkait pupuk untuk analisis lebih mendalam dan menerapkan metode klasterisasi lainnya untuk meningkatkan hasil evaluasi.

REFERENSI

1. Indah Jauhari and Dandy Kurnia, "Faktor yang Memengaruhi Keputusan Pembelian Produk Fashion Secara Online Melalui Aplikasi E-Commerce pada Generasi Milenial di Jakarta," *J. Ilm. Multidisiplin*, vol. 1, no. 2, pp. 09–18, 2022, doi: 10.56127/jukim.v1i2.90.
2. A. Ahdiat, "Tren Pengunjung E-Commerce Kuartal III 2023, Shopee Kian Melesat," *Databoks Katadata*, 2023. https://databoks.katadata.co.id/datapublish/2023/10/11/tren-pengunjung-e-commerce-kuartal-iii-2023-shopee-kian-melesat#:~:text=Menurut data SimilarWeb%2C 5 situs e-commerce kategori *marketplace*,2023 adalah Shopee%2C Tokopedia%2C Lazada%2C Blibli%2C dan Bu.
3. [BPS] Badan Pusat Statistik, "Berita Resmi Statistik Hasil Sensus Penduduk 2020," *Bps.Go.Id*, no. 27, pp. 1–52, 2020, [Online]. Available: https://papua.bps.go.id/pressrelease/2018/05/07/336/indeks-pembangunan-manusia-provinsi-papua-tahun-2017.html.
4. C. M. Annur, "E-Commerce Terpopuler di Kalangan Anak Muda, Siapa Juaranya?," *Databoks Katadata*, 2022. https://databoks.katadata.co.id/datapublish/2022/06/28/e-commerce-terpopuler-di-kalangan-anak-muda-siapa-juaranya (accessed Jun. 07, 2024).
5. Kementerian Pertanian, *Rencana Strategis Kementerian Pertanian Tahun 2020-2024*. 2021, pp. 1–161.
6. K. Chowdhury, D. Chaudhuri, and A. K. Pal, "An Entropy-based initialization method of *K-Means* clustering on the optimal number of clusters," *Neural Comput. Appl.*, vol. 33, no. 12, pp. 6965–6982, 2021, doi: 10.1007/s00521-020-05471-9.
7. Menteri Pertanian, "Peraturan Menteri Pertanian Nomor 04 Tahun 2019 tentang Pedoman Gerakan Pembangunan Sumber Daya Manusia Pertanian Menuju Lumbung Pangan Dunia 2045," pp. 1–12, 2019.
8. Badan Pusat Statistik, "Hasil Pencacahan Lengkap Sensus Pertanian 2023 - Tahap I," in *Badan Pusat Statistik*, 2023, pp. 1–343.
9. Y. H. Wulandari, Nina, "Determinants of Paddy Production in Indonesia : Study of Government Expenditures in Food Subsector and Climate Change," Program Studi Magister Perencanaan dan Kebijakan Publik Fakultas Ekonomi dan Bisnis UI, 2018.
10. D. L. Amaliah and N. F. Deli, "Internet, 'Pupuk' untuk Pertanian Masa Kini," *Bigdata.Bps.Go.Id*, pp. 1–10, 2023.
11. N. Sureja, B. Chawda, and A. Vasant, "An Improved *K-Medoids* Clustering Approach Based on The Crow Search Algorithm," *J. Comput. Math. Data Sci.*, vol. 3, no. April, p. 100034, 2022, doi: 10.1016/j.jcmds.2022.100034.
12. L. R. Singrapati, R. Dorab, and R. Kurniawan, "Pengelompokkan Toko Kaus Termurah E-Commerce Shopee berdasarkan Reputasi Toko Menggunakan Metode Clustering," *J. Sist. dan Teknol. Inf.*, vol. 12,

no. 1, pp. 65–72, 2024, doi: 10.26418/justin.v12i1.69067.
13. S. A. El-Khatib, Y. A. Skobtsov, and S. I. Rodzin, "Comparison of Hybrid ACO-*K-Means* Algorithm and Grub Cut for MRI Images Segmentation," *Procedia Comput. Sci.*, vol. 186, pp. 316–322, 2021, doi: 10.1016/j.procs.2021.04.150.
14. T. Mátrai and J. Tóth, "Cluster Analysis of Public Bike Sharing Systems for Categorization," *Sustain.*, vol. 12, no. 12, pp. 13–16, 2020, doi: 10.3390/SU12145501.
15. C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A Quantitative Discriminant Method of *Elbow* Point for The Optimal Number of Clusters in Clustering Algorithm," *Eurasip J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.
16. E. Garcia-Morato, M. J. Algar, C. Alfaro, F. Ortega, J. Gomez, and J. M. Moguerza, "Using *K-Medoids* for Distributed Approximate Similarity Search with Arbitrary Distances," 2024, [Online]. Available: http://arxiv.org/abs/2405.13795.
17. K. P. Sinaga and M. S. Yang, "Unsupervised *K-Means* clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
18. M. Shutaywi and N. N. Kachouie, "*Silhouette* Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021, doi: 10.3390/e23060759.
19. U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002, doi: 10.1109/TPAMI.2002.1114856.
20. M. A. Baig *et al.*, "Regression Analysis of Hydro-meteorological Variables for Climate Change Prediction: A Case Study of Chitral Basin, Hindukush Region," *Sci. Total Environ.*, vol. 793, p. 148595, 2021, doi: 10.1016/j.scitotenv.2021.148595.
21. T. Hennig-Thurau *et al.*, "The Impact of New Media on Customer Relationships," *J. Serv. Res.*, vol. 13, no. 3, pp. 311–330, 2010, doi: 10.1177/1094670510375460.
22. J. Gómez, B. Pérez-Aradros, and I. Salazar, "How to Beat Early Movers: The Role of Competitive Strategy and Industry Dynamism on Followers' Performance in The Telecommunications Industry," *Long Range Plann.*, vol. 55, no. 5, 2022, doi: 10.1016/j.lrp.2022.102244.
23. Z. Lv, Y. Jin, and J. Huang, "How Do Sellers Use Live *Chat* to Influence Consumer Purchase Decision in China?," *Electron. Commer. Res. Appl.*, vol. 28, pp. 102–113, 2018, doi: 10.1016/j.elerap.2018.01.003.
24. A. Guizzardi, L. V. Ballestra, and E. D'Innocenzo, "Hotel Dynamic Pricing, Stochastic Demand and Covid-19," *Ann. Tour. Res.*, vol. 97, p. 103495, 2022, doi: 10.1016/j.annals.2022.103495.