BAB III

ARTIKEL KARYA ILMIAH

PENERAPAN DATA MINING DALAM ANALISIS POLA PENJUALAN PRODUK CETAKAN DENGAN K-MEANS CLUSTERING

Oleh Reina Ahlaq Karim Narsa Sari[1], Deborah Kurniawati[2]* [1,2]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia
***Email:** [1]reina.ahlaq@students.utdi.ac.id, [2*]debbie@utdi.id

Article History:
Received: 14-04-2025
Revised: 20-04-2025
Accepted: 17-05-2025

Keywords:
K-Means, Clustering, Elbow Method, Silhouette Score, Pola Penjualan

Abstract: *Penelitian ini bertujuan untuk menganalisis pola penjualan produk cetakan dengan menerapkan metode data mining menggunakan algoritma K-Means Clustering. Data yang digunakan merupakan data penjualan produk cetakan dari wilayah Sulawesi, Indonesia. Tahapan penelitian meliputi exploratory data analysis (EDA), preprocessing data untuk normalisasi dan encoding, penerapan metode Elbow untuk menentukan jumlah cluster optimal, serta evaluasi hasil clustering menggunakan Silhouette Score. Hasil penelitian menunjukkan terbentuknya dua cluster optimal dengan nilai Silhouette Score sebesar 0.56. Masing-masing cluster memiliki karakteristik produk yang berbeda, di mana Cluster 1 memiliki keragaman produk yang tinggi, dengan beberapa jenis produk menonjol namun tidak dominan secara tunggal. Sementara itu, Cluster 0 cenderung terfokus pada sedikit produk utama yang mendominasi. Perbedaan ini menunjukkan adanya segmentasi pasar yang jelas, yang dapat dimanfaatkan dalam strategi pengelolaan stok, perencanaan produksi, dan pemasaran. Penelitian ini memberikan wawasan strategis bagi perusahaan dalam memahami preferensi konsumen berdasarkan distribusi produk dalam setiap cluster. Untuk pengembangan lebih lanjut, disarankan penggunaan algoritma alternatif seperti DBSCAN untuk mengatasi keterbatasan K-Means dalam mengidentifikasi cluster dengan distribusi data yang tidak beraturan.*

PENDAHULUAN

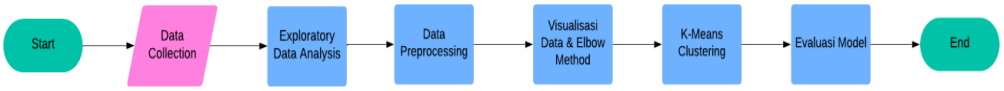
Di era digital yang berkembang pesat, dunia bisnis tidak hanya untuk bertahan melainkan harus terus bertumbuh melalui pengambilan keputusan yang cerdas dan berbasis data. Perkembangan teknologi informasi telah menciptakan peluang besar bagi perusahaan untuk menggali wawasan strategis dari data yang mereka miliki. Salah satu aset penting yang sering kali belum dimanfaatkan secara optimal adalah data penjualan. Melalui pendekatan data mining, informasi yang tersembunyi dalam data tersebut dapat diolah menjadi pengetahuan yang bernilai tinggi untuk mendukung arah strategi produksi, pemasaran, hingga pengembangan bisnis secara menyeluruh. Clustering merupakan teknik pengelompokan data berdasarkan kesamaan karakteristik yang bertujuan untuk menemukan pola tersembunyi dalam dataset besar. Dalam dunia bisnis, hasil clustering dapat dimanfaatkan untuk mengevaluasi performa produk, mengenali perilaku pelanggan, dan menyusun strategi pemasaran yang lebih personal dan terarah. Penelitian yang dilakukan oleh menunjukkan bahwa penerapan algoritma K-Means pada data penjualan Toko Kecantikan Putri berhasil mengelompokkan produk berdasarkan karakteristik pembelian pelanggan, menghasilkan cluster yang dapat digunakan untuk strategi pemasaran yang lebih efektif.

Algoritma K-Means menjadi pilihan populer dalam implementasi data mining karena kesederhanaannya, efisiensi dalam hal waktu, serta kemampuannya menangani data dalam skala besar. Metode ini telah diterapkan secara luas dalam berbagai penelitian mulai dari analisis penjualan produk makanan dan minuman, fashion, industri percetakan, hingga jasa ekspedisi. Sementara itu, penelitian oleh merancang aplikasi data mining untuk menentukan tingkat kelarisan produk menggunakan algoritma K-Means di Rima Market. Hasil clustering mampu membedakan produk yang laris dan tidak laris, sehingga membantu pihak swalayan dalam pengambilan keputusan untuk penambahan atau pengurangan stok secara tepat sasaran. Menurut penelitian pada unit Print on Demand Gramedia berhasil mengelompokkan produk cetakan berdasarkan volume transaksi yang kemudian digunakan sebagai dasar dalam penetapan strategi bisnis. Sementara itu penelitian oleh, mengimplementasikan clustering pada sektor fashion untuk memahami preferensi pelanggan, sehingga dapat menghindari penumpukan stok serta meningkatkan efisiensi distribusi produk.

Meskipun telah banyak penelitian yang memanfaatkan K-Means untuk mengelompokkan data penjualan, sebagian besar studi sebelumnya masih terbatas pada tahap segmentasi atau klasifikasi produk. Belum banyak yang menggali lebih dalam mengenai distribusi dan keragaman jenis produk dalam setiap cluster, maupun perbedaan signifikan antar cluster, yang padahal sangat penting dalam merumuskan strategi bisnis yang lebih terarah dan berdampak nyata. Berdasarkan hal tersebut, penelitian ini bertujuan untuk menganalisis distribusi dan keragaman jenis produk dalam setiap cluster, dengan menyajikan informasi berupa persentase dominansi produk dan perbedaan signifikan antar cluster. Hasil analisis ini diharapkan dapat memberikan wawasan yang komprehensif dalam mendukung strategi segmentasi pasar, perencanaan produksi, serta pengambilan keputusan yang lebih presisi dan berkelanjutan.

METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif, yaitu metode yang berfokus pada analisis data dalam bentuk angka (numerik) untuk mengidentifikasi pola tertentu. Dataset yang digunakan dalam penelitian ini berisi data penjualan produk cetakan pada tahun 2022 hingga 2023. Pada analisis clustering menggunakan dua variabel, yaitu jenis produk dan total penjualan, karena kedua variabel tersebut merepresentasikan pola penjualan produk. Berikut merupakan langkah-langkah penelitian yang dapat dilihat pada Gambar 1



Gambar 3.1 Flowchart Langkah Penelitian

1. Data Collection

Pada tahap awal data yang digunakan dalam penelitian ini merupakan data penjualan produk cetakan dari tahun 2022 hingga 2023. Dataset ini terdiri dari lima atribut utama, yaitu tanggal transaksi penjualan, jenis produk yang dijual, jumlah order yang menunjukkan banyaknya unit produk dalam satu transaksi, harga per unit dari produk yang dijual, dan total pendapatan yang diperoleh dari transaksi tersebut. Data ini dikumpulkan dalam format CSV dan digunakan sebagai dasar untuk analisis lebih lanjut dalam proses clustering.

2. Exploratory Data Analysis (EDA)

Pada tahap Exploratory Data Analysis, dilakukan analisis awal terhadap data untuk memahami karakteristik dan distribusi variabel. Proses ini melibatkan visualisasi distribusi data untuk melihat pola penjualan dari waktu ke waktu dan analisis tren penjualan untuk mengetahui produk dengan performa tinggi. Selain itu, pengecekan outlier juga dilakukan untuk mengidentifikasi nilai yang mungkin tidak sesuai atau ekstrem. Hasil dari Exploratory Data Analysis memberikan gambaran awal mengenai perilaku penjualan dan membantu dalam menentukan strategi preprocessing yang tepat. Exploratory Data Analysis memainkan peran penting dalam mengungkap pola awal secara efisien, sebagaimana ditekankan dalam penelitian oleh yang menyoroti bahwa proses eksplorasi awal dapat diarahkan ke titik-titik stabil dalam struktur data. Selain itu, mengintegrasikan PCA dalam tahap EDA untuk menyederhanakan variabel dan mempermudah 2 identifikasi pola sebelum pengelompokan dengan algoritma K-Means.

3. Data Preprocessing

Sebelum melakukan proses clustering, data perlu diproses agar lebih siap digunakan. Tahapan preprocessing meliputi penghapusan data yang tidak relevan atau duplikat untuk menghindari bias dalam model. Melakukan tahap normalisasi pada atribut numerik yaitu jumlah order, harga, dan total dengan menggunakan metode StandardScaler. Kemudian pada atribut kategori seperti jenis produk di encode menggunakan Label Encoding agar dapat diolah oleh algoritma K-Means. Penggunaan teknik encoding seperti ini juga telah diterapkan dalam penelitian oleh, di mana atribut kategorikal dikonversi menggunakan One Hot Encoding sebelum proses clustering dengan K-Means untuk memastikan format data sesuai dengan algoritma.

4. Elbow Method

Pada tahap ini, dilakukan visualisasi awal untuk melihat pola distribusi data sebelum clustering. Untuk menentukan jumlah cluster yang optimal menggunakan Metode Elbow. Berdasarkan penelitian yang dilakukan oleh metode Elbow digunakan untuk mengatasi kelemahan K-Means dalam menentukan jumlah cluster yang ideal. Metode ini bekerja dengan menghitung Sum of Squared Errors (SSE) pada berbagai jumlah cluster. Jumlah cluster yang optimal ditentukan pada titik di mana terjadi perubahan signifikan dalam grafik SSE yang membentuk titik siku (elbow point). Berikut merupakan rumus metode Elbow.

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 \quad (1)$$

Keterangan :
- k = jumlah cluster
- x_j = data ke-j dalam Cluster C_i
- μ_i= titik pusat cluster

Hasil dari metode Elbow akan digunakan untuk menentukan jumlah cluster terbaik yang akan digunakan dalam algoritma K-Means Clustering. Pada awalnya, centroid awal dipilih secara acak, kemudian setiap data akan dihitung jaraknya ke centroid menggunakan metode Euclidean Distance.

5. K-Means Clustering

Setelah menentukan jumlah cluster yang optimal menggunakan metode Elbow, tahap selanjutnya adalah penerapan algoritma K-Means Clustering. Algoritma ini bekerja dengan mengelompokkan data berdasarkan kemiripan karakteristiknya menggunakan pendekatan iteratif.

$$d(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} \quad (2)$$

Keterangan :

- $d(x, c)$ = jarak antara titik data x dan centroid c,
- x_i = nilai fitur ke-i dari data,
- c_i = nilai fitur ke-i dari centroid,
- n = jumlah fitur dalam dataset.

Meskipun algoritma K-Means dikenal karena kesederhanaan dan efisiensinya, berbagai penelitian terus mengembangkan variasinya untuk meningkatkan akurasi penelitian oleh Awad dan Hamad (2022), yang merancang implementasi K-Means berbasis neural engine terdistribusi untuk menangani big data secara lebih efisien.

6. Evaluasi Model (Silhouette Score)

Evaluasi dilakukan untuk menilai kualitas hasil clustering menggunakan Silhouette Score. Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan cluster yang baik. Berdasarkan penelitian yang dilakukan oleh, metode ini menghitung seberapa baik data berada dalam cluster tertentu dibandingkan dengan cluster lain. Evaluasi ini memastikan bahwa hasil clustering memberikan segmentasi yang bermakna bagi analisis lebih lanjut. Berikut merupakan rumus Silhouette Score.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

Keterangan:
- $a(i)$ = rata-rata jarak antara data i dan anggota klaster yang sama
- $b(i)$ = rata-rata jarak antara data i dan klaster terdekat
- Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai lebih tinggi menunjukkan clustering yang lebih baik.

HASIL DAN PEMBAHASAN

1. Data Collection

Data yang digunakan dalam penelitian ini merupakan data historis penjualan produk cetakan yang diperoleh dari salah satu perusahaan percetakan yang beroperasi di wilayah Sulawesi, Indonesia. Dataset mencakup periode waktu dari tahun 2022-2023. Secara keseluruhan data terdiri dari 1.076 baris data dengan lima atribut yaitu, tanggal transaksi penjualan, jenis produk, jumlah order, harga dan total pendapatan. Data ini kemudian digunakan sebagai dasar untuk proses preprocessing dan analisis lebih lanjut, khususnya dalam tahapan clustering menggunakan algoritma K-Means. Sebelum digunakan, dilakukan proses validasi awal untuk memastikan tidak ada nilai kosong (missing value) atau duplikasi data. Dataset ini bersifat kategori dan numerik, sehingga diperlukan transformasi seperti encoding dan normalisasi pada tahap preprocessing agar dapat digunakan dalam proses clustering. Seluruh proses pengolahan data dilakukan menggunakan tools seperti Python (pandas, matplotlib, seaborn, sklearn).

2. Exploratory Data Analysis (EDA)

Tahapan Exploratory Data Analysis (EDA) dilakukan untuk memahami karakteristik awal dari dataset penjualan yang digunakan. Melalui proses ini, dilakukan pengecekan struktur data, deteksi terhadap data duplikat dan data kosong, serta analisis distribusi data, khususnya pada atribut jenis produk. Analisis ini bertujuan untuk mendapatkan gambaran umum mengenai pola penjualan dan kondisi data sebelum dilakukan proses preprocessing dan clustering.

a. Frekuensi Kemunculan Produk dalam Transaksi

Analisis awal dilakukan dengan menghitung frekuensi kemunculan setiap jenis produk dalam data transaksi penjualan. Tujuan dari analisis ini adalah untuk mengidentifikasi produk mana yang paling sering muncul dalam transaksi selama periode tahun 2022 hingga 2023. Hasil perhitungan menunjukkan bahwa terdapat variasi frekuensi penjualan terhadap setiap produk, di mana beberapa produk lebih sering muncul dibandingkan produk lainnya. Hasil perhitungan ditampilkan dalam Tabel 1 berikut :

Tabel 1. Frekuensi Kemunculan Produk Teratas dalam Transaksi Penjualan (2022–2023)

Jenis Produk	Jumlah
Dupleks310	162
Ivory230	133
CraftLaminasi290	121
Dupleks350	78
GreaseProof	78
Ivory270	58
FoodpakMatte245	54
Ivory250	51
GraftFoodpak290	28
Ivory300	25

Berdasarkan Tabel 1, dapat disimpulkan bahwa produk Dupleks310, Ivory230, dan CraftLaminasi290 merupakan produk yang paling sering muncul dalam transaksi, sementara beberapa produk lain memiliki frekuensi kemunculan yang jauh lebih rendah.

b. Pengecekan dan Penghapusan Data Duplikat

Langkah selanjutnya dalam EDA adalah melakukan pengecekan terhadap adanya data duplikat dalam dataset. Duplikat dapat terjadi apabila terdapat transaksi yang tercatat lebih dari satu kali dengan informasi yang sama, dan hal ini dapat memengaruhi akurasi analisis. Pengecekan dilakukan menggunakan fungsi df.duplicated() untuk mengidentifikasi baris yang memiliki data identik. Hasil pengecekan menunjukkan bahwa terdapat beberapa data duplikat dalam dataset, seperti ditunjukkan pada Gambar 2.

```
#ngecek isi duplikatnya ada apa aja
df[df.duplicated()]
```

	Tanggal	Jenis Produk	Jumlah Order	Harga	Total
10	9/8/2022	Dupleks350	500	1800	900000
15	11/8/2022	Dupleks350	500	1800	900000
67	7/9/2022	Dupleks310	2000	1600	3200000
82	14/09/2022	Foodpak295	1000	1600	1600000
95	19/09/2022	Dupleks310	1000	1400	1400000

Gambar 3.2 Contoh Data Duplikat yang Ditemukan dalam Dataset

Setelah data duplikat berhasil diidentifikasi, langkah selanjutnya adalah menghapus data tersebut menggunakan fungsi df.drop_duplicated( ).

```
#ngecek lagi apakah duplikatnya sudah di hapus atau belum
df.duplicated().sum()
print('Jumlah duplikat setelah di drop:', df.duplicated().sum())
```

```
Jumlah duplikat setelah di drop: 0
```

Gambar 3.3 Jumlah Data Duplikat Setelah Proses Penghapusan

Gambar 3 menunjukkan bahwa penghapusan berhasil, dilakukan pengecekan ulang menggunakan fungsi df. duplicated( ). Sum( ) untuk menghitung jumlah duplikat yang tersisa. Hasilnya menunjukkan bahwa seluruh data duplikat berhasil dihapus.

c. Cek dan Penanganan Data Kosong (Missing Value)

Setelah menangani data duplikat, langkah selanjutnya adalah melakukan pengecekan terhadap data kosong (missing value). Hal ini penting untuk memastikan bahwa tidak ada nilai kosong pada atribut-atribut penting yang dapat mengganggu proses analisis maupun hasil clustering. Pengecekan dilakukan menggunakan fungsi df.isna().sum() untuk mengetahui jumlah nilai kosong pada setiap kolom. Berdasarkan hasil pengecekan, tidak ditemukan nilai kosong pada seluruh atribut yang tersedia. Oleh karena itu, tidak diperlukan proses penghapusan data kosong, dan data dapat langsung digunakan untuk tahapan analisis selanjutnya.

d. Statistik Deskriptif

Setelah data dipastikan bersih dari duplikat dan nilai kosong, dilakukan analisis deskriptif untuk melihat gambaran umum dari variabel numerik yang digunakan dalam penelitian. Statistik ini mencakup nilai minimum, maksimum, rata-rata (mean), standar

deviasi, serta kuartil dari masing-masing variabel seperti jumlah order, harga per unit, dan total pendapatan. Analisis ini membantu dalam memahami skala, variasi, serta distribusi data numerik sebelum dilakukan proses normalisasi. Hasilnya diperoleh menggunakan fungsi df.describe( ), yang memberikan ringkasan statistik dari setiap variabel numerik.

Tabel 2. Statistik Deskriptif Data

	Jumlah Order	Harga	Total
count	1036.000000	1036.000000	1.036×10^3
mean	1939.872587	1699.154440	2.372408×10^6
std	2540.335312	3414.781832	1.749042×10^6
min	5.000000	100.000000	1.537600×10^4
25%	1000.000000	950.000000	1.312500×10^6
50%	1000.000000	1500.000000	1.850000×10^6
75%	2000.000000	1800.000000	3.000000×10^6
max	40000.000000	99970.000000	2.149355×10^7

Tabel 2 menunjukkan ringkasan statistik dari 1.036 data transaksi. Pada kolom Jumlah Order, rata-rata jumlah unit produk dalam satu transaksi adalah sekitar 1.939 unit, dengan jumlah minimum 5 unit dan maksimum 40.000 unit, menunjukkan rentang kuantitas pemesanan yang sangat luas. Standar deviasi sebesar 2.540,34 menunjukkan penyebaran data yang cukup tinggi terhadap rata-rata. Untuk variabel Harga, nilai rata-rata berada di angka Rp 1.699,15 dengan harga terendah Rp 100 dan tertinggi mencapai Rp 99.970 per unit. Standar deviasi sebesar Rp 3.141,78 menunjukkan terdapat produk dengan harga yang sangat bervariasi, mencerminkan heterogenitas produk dalam dataset. Sementara itu, pada variabel Total Pendapatan, rata-rata pendapatan per transaksi sebesar Rp 2.372.408, dengan nilai minimum Rp 153.760 dan maksimum mencapai Rp 21.493.550. Nilai standar deviasi sebesar Rp 1.749.042 juga mencerminkan variasi besar antar transaksi, baik dari sisi jumlah produk yang dibeli maupun harganya, Informasi statistik deskriptif ini penting dalam tahap awal untuk memahami skala dan sebaran data sebelum dilakukan proses normalisasi dan analisis clustering lebih lanjut.

3. Data Preprocessing

Sebelum melakukan proses clustering, data perlu diproses agar siap digunakan dalam pemodelan. Tahap ini dilakukan untuk memastikan data berada dalam skala yang seragam serta dapat dikenali oleh algoritma yang digunakan. Pada tahap ini, dilakukan dua proses utama. Pertama, dilakukan encoding pada atribut jenis produk menggunakan Label Encoding, agar data kategori tersebut dapat diubah menjadi bentuk numerik yang bisa dibaca oleh algoritma K-Means. Kedua, seluruh atribut numerik termasuk hasil encoding tersebut dinormalisasi menggunakan StandardScaler. Proses ini menghasilkan distribusi data dengan rata- rata nol dan standar deviasi satu, yang penting untuk menghindari bias akibat perbedaan skala antar fitur.

Tabel 3. Contoh Data Product Encoded

	Jenis Produk	Product_Encoded
0	Foodpak260	40
1	FoodpakMatte245	49
2	CraftLaminasi290	21
3	CraftLaminasi290	21
4	Dupleks310	29
5	Dupleks310	29
6	Ivory270	72
7	Kinstruk130	79
8	HVS	59
9	Dupleks350	33

Tabel 3 menampilkan hasil proses encoding pada kolom kategorikal yaitu Jenis Produk, yang telah diubah ke dalam format numerik agar dapat dibaca oleh algoritma K-Means.

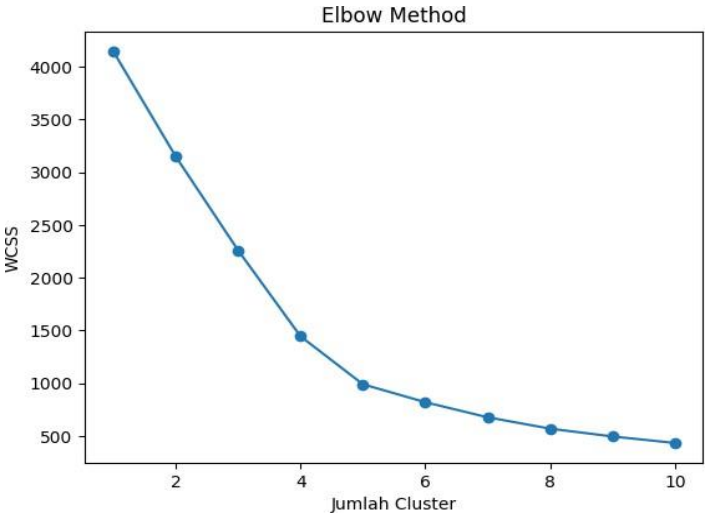
Tabel 4. Hasil Normalisasi

Tanggal	Jenis Produk	Jumlah Order	Harga	Total	Product_Encoded
05/08/2022	Foodpak260	-0.370158	0.029546	-0.327428	-0.212871
05/08/2022	FoodpakMatte245	-0.370158	0.058845	-0.270226	0.226344
05/08/2022	CraftLaminasi290	1.205197	-0.278089	0.788007	-1.140103
05/08/2022	CraftLaminasi290	-0.370158	-0.146245	-0.670638	-1.140103
07/08/2022	Dupleks310	-0.370158	-0.043700	-0.470432	-0.749690

Tabel 4 menunjukkan hasil normalisasi terhadap fitur numerik menggunakan metode StandardScaler, di mana data telah dikonversi ke skala standar dengan rata-rata 0 dan deviasi standar 1.

4. Penentuan Jumlah Cluster (Elbow Method)

Penentuan jumlah cluster yang optimal dalam penelitian ini dilakukan menggunakan pendekatan kombinasi antara Elbow Method dan Silhouette Score. Pada awalnya, dilakukan visualisasi Elbow Method dengan menghitung nilai Sum of Squared Errors (SSE) untuk beberapa nilai k (jumlah cluster), sebagaimana ditunjukkan pada Gambar 4. Namun, grafik yang dihasilkan tidak menunjukkan titik siku (elbow) yang jelas sebagai indikasi jumlah cluster optimal. Oleh karena itu, digunakan metrik tambahan yaitu Silhouette Score untuk membantu proses pengambilan keputusan. Silhouette Score mengukur seberapa baik data dikelompokkan dalam sebuah cluster, dengan mendekati nilai 1 menunjukkan pemisahan cluster yang optimal. Dari hasil evaluasi, nilai Silhouette Score tertinggi sebesar 0,56 ditemukan pada saat jumlah cluster $k = 2$. Berdasarkan hal tersebut, maka diputuskan untuk menggunakan 2 cluster dalam proses K-Means clustering.



Gambar 3.4 Grafik Elbow Method

Oleh karena itu, digunakan metrik tambahan yaitu Silhouette Score untuk membantu proses pengambilan keputusan. Silhouette Score mengukur seberapa baik data dikelompokkan dalam sebuah cluster, dengan nilai mendekati 1 menunjukkan pemisahan cluster yang optimal. Tabel 5 menunjukkan nilai Silhouette Score untuk berbagai nilai *k*.

Tabel 5. Hasil Evaluasi Silhouette Score untuk Berbagai Jumlah Cluster (K)

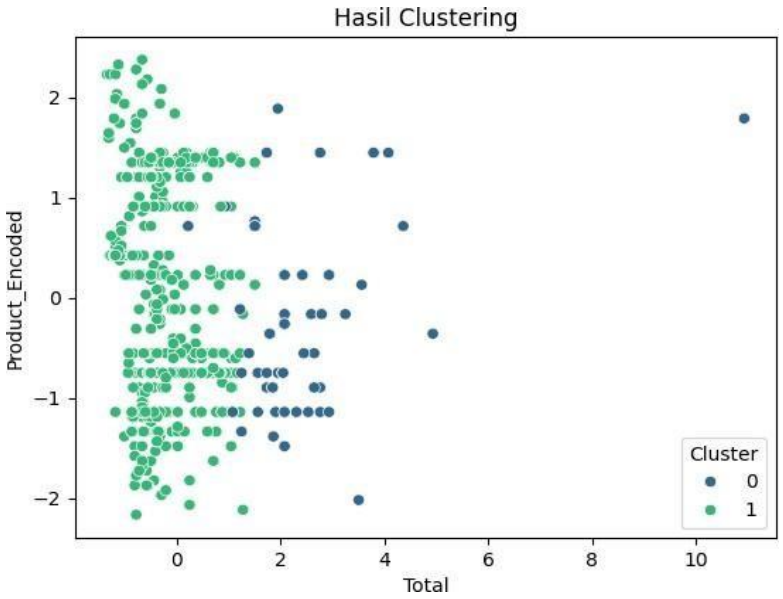
K	Silhouette Score
2	0.564510
3	0.401517
4	0.462490
5	0.481748
6	0.497878
7	0.457055
8	0.397586
9	0.400031
10	0.427038

Tabel 5 menunjukkan hasil evaluasi tersebut, nilai Silhouette Score tertinggi sebesar 0,5645 ditemukan saat jumlah cluster $k = 2$. Oleh karena itu, ditetapkan bahwa pemodelan K-Means clustering akan menggunakan 2 cluster.

5. Clustering dengan K-Means

Pada tahap ini, dilakukan proses pengelompokan data menggunakan algoritma K-Means. K-Means merupakan metode clustering yang membagi data ke dalam sejumlah kelompok (cluster) berdasarkan kemiripan karateristiknya. Pada penelitian ini jumlah cluster yang digunakan adalah k=2, yang ditentukan berdasarkan hasil evaluasi menggunakan Silhouette Score. Langkah awal dari algoritma K-Means dimulai dengan memilih centroid (titik pusat cluster) secara acak dari data yang tersedia. Centroid ini

berfungsi sebagai acuan untuk mengelompokkan data. Selanjutnya, setiap data akan dihitung jaraknya ke masing-masing centroid, kemudian dikelompokkan ke dalam cluster yang memiliki jarak terdekat. Setelah semua data data terbagi ke dalam cluster. Proses ini dilakukan secara berulang hingga data tidak lagi berpindah cluster, atau sampai posisi centroid sudah tidak berubah secara signifikan (konvergen). Hasil dari proses clustering ini memberikan dua kelompok utama berdasarkan pola penjualan produk, yang dapat digunakan untuk memahami karakteristik masing-masing cluster, seperti perbedaan ratarata jumlah order atau total pendapatan.



Gambar 3.5 Visualisasi hasil clustering K-Means

Gambar 5 menunjukkan hasil evaluasi dua cluster yang terbentuk berdasarkan atribut Total dan *Product_Encode*. Terlihat bahwa data terbagi ke dalam dua kelompok yang berbeda, dengan pola penyebaran yang cukup jelas. Cluster ditandai dengan warna yang berbeda, yaitu hijau untuk Cluster 1 dan biru untuk Cluster 0.

Tabel 6. Distribusi Jumlah Data per Cluster

Cluster	Jumlah Data
1	934
0	102

Berdasarkan Tabel 6, diketahui bahwa Cluster 1 memiliki jumlah data yang lebih dominan dibandingkan dengan Cluster 0, yaitu sebanyak 934 data. Hal ini menunjukkan bahwa sebagian besar data penjualan memiliki karakteristik yang mirip dengan Cluster 1.

Tabel 7. Statistik deskriptif untuk Cluster 1

	Cluster	Jumlah Order	Harga	Total
count	934.0	934.000000	934.000000	9.340000e+02
mean	1.0	1916.368308	1614.592077	2.321241e+06
std	0.0	2603.249616	1536.924877	1.621044e+06
min	1.0	5.000000	100.000000	1.537600e+04
25%	1.0	1000.000000	950.000000	1.300000e+06
50%	1.0	1000.000000	1500.000000	1.800000e+06
75%	1.0	2000.000000	1800.000000	3.000000e+06
max	1.0	40000.000000	18000.000000	1.100000e+07

Berdasarkan Tabel 7, Cluster 1 merupakan cluster dengan jumlah data terbanyak, yaitu 934 data. Hal ini menunjukkan bahwa sebagian besar data penjualan memiliki karakteristik yang serupa dengan cluster ini. Rata-rata jumlah order per transaksi di Cluster adalah 1.916 unit, dengan harga rata-rata produk sebesar Rp. 1.614 dan total penjualan per transaksi sebesar Rp 2.321.241. Standar deviasi yang cukup tinggi pada variabel Jumlah order ±2603 dan Total Penjualan ±1.6 juta menunjukkan adanya variasi yang cukup besar dalam volume dan nilai penjualan di cluster ini. Meskipun demikian, nilai kuartil menunjukkan bahwa sebagian besar transaksi masih berada pada rentang nilai yang cukup stabil. Karakteristik ini menggambarkan cluster yang berisi produk-produk dengan harga dan volume penjualan yang bervariasi, tetapi cenderung berada pada kisaran menengah ke bawah.

Tabel 8. Statistik deksriptif untuk Cluster 0

	Cluster	Jumlah Order	Harga	Total
count	102.0	102.000000	102.000000	1.020000e+02
mean	0.0	1910.000000	2543.039216	2.521898e+06
std	0.0	1855.417676	9854.937627	2.532950e+06
min	0.0	5.000000	250.000000	7.500000e+04
25%	0.0	1000.000000	950.000000	1.300000e+06
50%	0.0	1000.000000	1500.000000	1.800000e+06
75%	0.0	2000.000000	1800.000000	2.962500e+06
max	0.0	10000.000000	99970.000000	2.149355e+07

Pada Tabel 8 menunjukkan bahwa Cluster 0 terdiri dari 102 data, jauh lebih sedikit dibandingkan dengan Cluster 1. Namun, rata-rata harga produk di cluster ini lebih tinggi, yaitu sekitar Rp 2.543 per unit, dengan total penjualan rata-rata mencapai Rp. 2.518.898 per transaksi. Standar deviasi yang tinggi pada variabel Harga ±9854 dan Total Penjualan

±2.5 juta menunjukkan ketimpangan atau variasi yang cukup besar antar data dalam cluster ini. Hal ini menunjukkan bahwa Cluster 0 didominasi oleh produk-produk dengan nilai jual tinggi, meskipun frekuensinya tidak sebanyak Cluster 1. Karakteristik ini bisa diasosiasikan dengan segmen pasar khusus atau produk premium.

6. Evaluasi Model

Setelah proses clustering selesai dilakukan, langkah selanjutnya adalah mengevaluasi hasil cluster yang terbentuk. Evaluasi dilakukan menggunakan metode Silhouette Score, yaitu metrik yang mengukur seberapa baik objek berada dalam clusternya masing-masing dibandingkan dengan cluster lainnya. Nilai Silhouette Score berkisar antara -1 hingga 1. Semakin mendekati 1 maka semakin baik kualitas clustering tersebut. Sedangkan nilai yang mendekati nilai 0 menunjukkan bahwa data berada di antara dua cluster dan nilai negatif mengindikasikan bahwa data kemungkinan salah pengelompokkannya. Pada penelitian ini, evaluasi dilakukan terhadap beberapa nilai k (jumlah cluster), dan hasil menunjukkan bahwa nilai Silhouette Score tertinggi berada pada k=2 dengan skor sebesar 0.56. Hal ini menunjukkan bahwa pembagian cluster menjadi dua kelompok memberikan segmentasi yang paling baik dan bermakna dibandingkan jumlah cluster lainnya. Oleh karena itu, jumlah cluster optimal yang digunakan dalam penelitian ini adalah 2 cluster.

KESIMPULAN

Berdasarkan hasil clustering yang telah dilakukan, dapat disimpulkan bahwa masing-masing cluster memiliki karakteristik produk yang berbeda. Distribusi jenis produk menunjukkan perbedaan yang jelas antara Cluster 0 dan Cluster 1. Cluster 1 memiliki keragaman produk yang tinggi, dengan *Dupleks310* (15,20%), *Ivory230*(13,38%), *CraftLaminasi290* (9,85%), dan *GreaseProof* (8,03%) sebagai produk-produk yang relatif dominan. Namun, banyak produk lain di Cluster 1 memiliki persentase yang rendah. Cluster 0 sebaliknya lebih terfokus dengan *CraftLaminasi290* sebagai produk utama (27,45%), diikuti oleh *Foodpak260Glossy* dan *HVS* (masing-masing 10,78%). Perbedaan ini mengisyaratkan strategi bisnis yang berbeda, di mana Cluster 1 mungkin melayani pasar yang lebih luas dan Cluster 0 lebih terspesialisasi. Temuan ini memberikan wawasan penting bagi perusahaan dalam memahami segmentasi pasar berdasarkan karakteristik produk. Dengan mengenali pola distribusi produk di setiap cluster, perusahaan dapat menyusun strategi produksi dan pemasaran yang lebih tepat sasaran, misalnya dengan mengalokasikan stok secara lebih efisien atau menyesuaikan promosi dengan preferensi konsumen di masing-masing cluster. Di outlier atau bentuk distribusi tidak beraturan, penelitian selanjutnya dapat mempertimbangankan penggunaan algoritma alternatif seperti DBSCAN, yang terbukti memiliki validitas klaster yang lebih baik pada jenis data serupa

Ucapan Terima kasih

Saya sebagai penulis mengucapkan terima kasih kepada Ibu Deborah Kurniawati atas bimbingan dan arahannya dalam penyusunan penelitian ini. Terima kasih untuk kedua orang tua saya yang selalu memberikan dukungan selama proses penelitian berlangsung, serta kepada teman-teman saya yang selalu memberikan semangat, motivasi dalam menyelesaikan penelitian ini. Saya juga mengucapkan terima kasih kepada semua pihak yang telah berkontribusi, baik secara langsung maupun tidak langsung, dalam mendukung kelancaran penelitian ini. Semoga penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan menjadi referensi yang berguna bagi penelitian selanjutnya.