JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

Analisis Segmentasi Pelanggan Mall Menggunakan Metode K-Means Clustering Dan Elbow Untuk Optimasi Strategi Pemasaran

Yokshane Adryan[1], Fathan Rizqi[2], Ivan Bayu[3], Rio Yosafat[4], Saprudi[5*]

[1,2,3,4,5]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, UniversitasPamulang, Kota, Indonesia
Email: [1]andreadryan130@gmail.com, [2]fathanrizqi029@gmail.com, [5*]dosen00845@unpam.ac.id
(* : coressponding author)

Abstrak - Pemahaman mendalam tentang pelanggan cukup penting bagi perusahaan dalam menghadapi persaingan pasar yang ketat. Segmentasi pelanggan menjadi salah satu strategi utama untuk memahami dan mengelompokkan pelanggan berdasarkan karakteristik dan perilaku tertentu, seperti demografi, pendapatan, serta pola belanja. Penelitian pada jurnal ini menggunakan algoritma K-Means Clustering yang didukung metode elbow untuk dapat menentukan klaster optimal dalam dataset pelanggan mall. Proses analisis mencakup eksplorasi data, pre-processing, dan transformasi data untuk memastikan kualitas dataset sebelum diterapkan metode clustering. Dataset Mall_Customers dengan lima atribut utama—usia, jenis kelamin, pendapatan tahunan, skor pengeluaran, dan ID pelanggan—digunakan untuk mengidentifikasi empat klaster utama dengan karakteristik yang spesifik. Klaster yang dihasilkan menunjukkan pola belanja unik yang mencakup pelanggan dengan preferensi terhadap produk premium hingga pelanggan dengan pembelian kuantitas tinggi tetapi ekonomis. Penelitian ini memberikan wawasan yang bermanfaat bagi perusahaan dalam menyusun strategi pemasaran yang lebih efektif dan terarah berdasarkan klaster pelanggan. Penggunaan metode K-Means dan elbow terbukti memberikan hasil yang signifikan dalam analisis segmentasi pelanggan berbasis data.

Kata Kunci: Segmentasi Pelanggan; Algoritma K-Means; Metode Elbow; Strategi Pemasaran; Analisis Data

Abstract *- A deep understanding of customers is essential for companies in facing intense market competition. Customer segmentation is one of the main strategies to understand and group customers based on certain characteristics and behaviors, such as demographics, income, and shopping patterns. This study uses the K-Means Clustering algorithm supported by the elbow method to determine the optimal number of clusters in the mall customer dataset. The analysis process includes data exploration, pre-processing, and data transformation to ensure the quality of the dataset before applying the clustering method. The Mall_Customers dataset with five main attributes—age, gender, annual income, spending score, and customer ID—is used to identify four main clusters with specific characteristics. The resulting clusters show unique shopping patterns that include customers with a preference for premium products to customers with high-quantity but economical purchases. This study provides useful insights for companies in developing more effective and targeted marketing strategies based on customer clusters. The use of the K-Means and elbow methods has been proven to provide significant results in data-based customer segmentation analysis.*

Keywords*: Customer Segmentation; K-Means Algorithm; Elbow Method; Marketing Strategy; Data Analysis*

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat telah mendorong perubahan signifikan di berbagai bidang, termasuk bisnis. Dalam dunia bisnis, persaingan yang ketat menuntut perusahaan untuk memahami karakteristik pelanggan guna menciptakan strategi yang tepat dan mempertahankan posisi di pasar. Salah satu sektor yang mengalami transformasi besar adalah e-commerce, yang memnunjukan perubahan dinamis dalam perilaku suatu pembelian pelanggan. Fokus utama perusahaan kini tidak hanya pada produk, tetapi juga pada pengalaman pelanggan (Customer Experience) untuk meningkatkan loyalitas dan mempertahankan basis pelanggan yang ada.

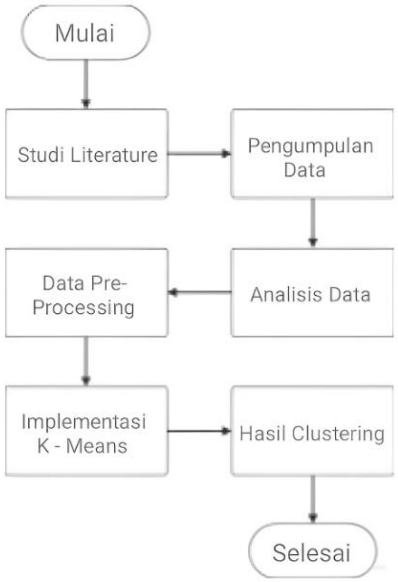
Salah satu strategi penting dalam memahami pelanggan adalah segmentasi, yaitu proses membagi pelanggan ke dalam kelompok-kelompok berdasarkan karakteristik tertentu seperti demografi, kebiasaan membeli, dan nilai transaksi. Segmentasi ini menjadi kunci dalam merancang promosi dan strategi pemasaran yang relevan untuk setiap segmen pelanggan. Namun, tantangan besar muncul ketika volume data yang besar harus dianalisis untuk mendapatkan segmentasi yang optimal. Oleh karena itu, dibutuhkan pendekatan berbasis data seperti teknik data mining dan algoritma clustering.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

Penelitian ini berfokus pada analisis segmentasi pelanggan menggunakan dataset *Mall Customers*, dengan tujuan memahami perilaku pelanggan berdasarkan atribut seperti usia, pendapatan tahunan, dan pengeluaran tahunan. Metode yang digunakan mengadopsi pendekatan clustering untuk membagi pelanggan ke dalam beberapa klaster yang relevan. Hasil penelitian ini diharapkan dapat memberikan wawasan yang berguna bagi perusahaan dalam mengembangkan strategi pemasaran yang lebih terarah dan efektif.

2. METODE

Bagian ini menjelaskan alur dari metode penelitian yang digunakan.Berikut Flowchart alur metode penelitian akan disajikan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

2.1 Studi Literature

Pada tahap ini, dilakukan kajian terhadap jurnal dan penelitian terdahulu yang membahas metode segmentasi pelanggan. Studi literatur ini bertujuan untuk mengidentifikasi metode yang paling sesuai dan efektif, terutama dalam penerapan algoritma clustering seperti K-Means untuk segmentasi pelanggan.

2.2 Pengumpulan Data

Tahap ini melibatkan pengumpulan data yang relevan dengan penelitian. Dataset yang digunakan adalah Mall Customers, yang berisi informasi terkait pelanggan, seperti Customer ID, jenis kelamin, usia, pendapatan tahunan, dan skor pengeluaran. Dataset ini terdiri dari 200 data dengan 5 atribut.

2.3 Analisis Data

Langkah ini mencakup eksplorasi data untuk memahami karakteristik dataset, seperti distribusi variabel, hubungan antar atribut, dan pola-pola yang relevan. Analisis ini memberikan gambaran awal mengenai data yang akan digunakan dalam segmentasi pelanggan.

2.4 Data Pre-Processing

Tahap ini bertujuan untuk mempersiapkan data agar siap digunakan dalam proses clustering. Dengan menggunakan dataset *Mall Customers*, langkah-langkah yang dilakukan adalah:

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

a. Mengatasi Nilai Kosong (Missing Values): Memastikan tidak ada data kosong. Jika ditemukan, data tersebut dapat dihapus atau diisi dengan nilai rata-rata, median, atau modus sesuai atributnya.

b. Transformasi Data: Menerapkan normalisasi atau *feature scaling* untuk atribut numerik seperti pendapatan tahunan dan skor pengeluaran, guna memastikan data berada pada skala yang sama.

c. Seleksi Data: Memilih atribut yang relevan untuk clustering, seperti usia, pendapatan tahunan, dan skor pengeluaran.

2.5 Implementasi K-Means Clustering

Pada tahap ini, algoritma K-Means diterapkan untuk mengelompokkan pelanggan ke dalam beberapa klaster. Penentuan jumlah klaster dilakukan dengan menggunakan metode *elbow*, yang mengevaluasi nilai K terbaik berdasarkan variasi intra-klaster. Setelah klaster ditentukan, pelanggan dikelompokkan berdasarkan karakteristik yang serupa.

2.6 Hasil Clustering

Hasil akhir berupa segmentasi pelanggan yang memisahkan beberapa pelanggan ke dalam suatu kelompok berdasarkan perilaku dan karakteristik pembelian mereka. Setiap klaster dianalisis untuk memahami kontribusinya terhadap bisnis, yang kemudian dapat digunakan untuk menyusun strategi pemasaran yang lebih efektif.

3. ANALISA DAN PEMBAHASAN

Pada bagian ini berisikan hasil penelitian berdasarkan metode yang sudah ditentukan sebelumnya. Berikut adalah hasil dan pembahasan dari pada penelitian yang dilakukan.

3.1 Studi Literature

Diawali dengan proses studi literatur yang bertujuan untuk menemukan referensi relevan sesuai dengan objek-objek penelitian yang ada. Proses penelusuran sumber dilakukan melalui database seperti Google Scholar, IEEE Explore, dan banyak lagi, dengan kriteria sumber dapat diakses secara terbuka (open access) dan dapat diunduh. Untuk mempermudah pencarian, digunakan alat bantu "Publish or Perish" dengan memasukkan kata kunci yang relevan, misalnya “Customer Segmentation” + “K-means”. Pencarian ini diarahkan untuk mengidentifikasi metode yang cukup banyak digunakan dalam menyelesaikan permasalahan segmentasi pada pelanggan guna memperoleh hasil yang optimal.

Sebanyak 5 artikel jurnal yang memenuhi kriteria telah dipilih berdasarkan rentang waktu lima tahun terakhir, yaitu dari 2020 hingga 2022. Jurnal-jurnal tersebut membahas topik terkait segmentasi pada pelanggan menggunakan algoritma k-means dan *method* elbow . Jurnal yang diperoleh kemudian dikelompokkan dalam tabel 1.

Tabel.1 Hasil Jurnal Terpilih Sesuai Dengan Topik

Cites	Penulis	Variable Jurnal Penelitian	Sumber Jurnal Penelitian	Jumlah	Tahun
[25]	ABH Kiat	K-Means Algorithm, Metode Elbow, RFM Model	Jurnal Sistem Informasi (Journal of Information Systems)	1	2020
[24]	R Taghi Livari, N	RFM Model, K-Means Algorithm, Food Distribution Industry	Jurnal Repositor	1	2021
[23]	P. Anitha	RFM model, customer purchase behavior,K-Means Algorithm	SRPH (Scientific ResearchPublishing House)	1	2022

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

[21]	T Juhari, A Juarna	RFM Model, K-Means Algorithm, Online Bookstore	Proceedings - 2018 IEEE 15thInternational Conference on e-Business Engineering, ICEBE 2018	1	2022
[20]	A Alamsyah,PE Prasetyo,S Sunyoto	RFM Model, K-Means Algorithm, Elbow Method	Mathematical Problems in Engineering	1	2022

Berdasarkan tabel 1, analisis dari berbagai literatur menunjukkan bahwa algoritma k-means dan *method* elbow tetap menjadi salah satu pendekatan dalam analisis data. Dalam rentang waktu 2020–2022, metode ini banyak digunakan untuk segmentasi data, yaitu proses pembagian data ke dalam suatu kelompok yang lebih terdefinisi. Beberapa penelitian juga memanfaatkan *method* elbow untuk menentukan jumlah klaster optimal. *Method* elbow membantu dalam mengidentifikasi jumlah klaster yang sesuai dengan data yang dianalisis.

Analisis berbagai sumber menunjukkan bahwa algoritma k-means dapat memberikan hasil yang baik dan andal, termasuk segmentasi data. Metode elbow turut mendukung peneliti untuk menentukan parameter seperti jumlah klaster optimal untuk meningkatkan efektivitas analisis data. Dengan demikian, kombinasi metode k-means dan *method* elbow terbukti memberikan kontribusi yang lumayan signifikan dalam analisis data serta pengelompokan data secara keseluruhan.

3.2 Pengumpulan Data

Penelitian ini menggunakan dataset *Mall_Customers.csv* yang berisi data pelanggan dari sebuah pusat perbelanjaan. Dataset tersebut memiliki 5 kolom yang mencakup informasi demografis dan perilaku pelanggan. Tabel berikut menjelaskan setiap atribut yang terdapat dalam dataset:

Tabel.2 Deskripsi Dataset

Kategori	Keterangan
CustomerID	ID unik untuk mengidentifikasi masing-masing pelanggan
Gender	Jenis kelamin pelanggan (pria atau wanita)
Age	Usia pelanggan dalam satuan tahun
Annual Income	Pendapatan tahunan pelanggan dalam ribuan dolar
Spending Score	Skor yang mencerminkan pola pengeluaran pelanggan, dinilai antara 1 hingga 100

Dataset ini dipilih karena relevan dengan tujuan penelitian, yaitu melakukan segmentasi pelanggan berdasarkan data yang tersedia.

3.3 Analisis Data

Langkah ini bertujuan untuk mengeksplorasi dataset **Mall_Customers** guna memahami karakteristik pelanggan yang akan digunakan dalam segmentasi. Analisis distribusi variabel menunjukkan bahwa pelanggan memiliki rentang usia antara 18 hingga 70 tahun dengan rata-rata 38,85 tahun, sedangkan pendapatan tahunan berkisar antara 15 hingga 137 ribu dolar dengan rata-rata 60,56 ribu dolar. Skor pengeluaran pelanggan, yang mencerminkan tingkat aktivitas belanja, memiliki rata-rata sebesar 50,20 dengan nilai minimum 1 dan maksimum 100. Hasil ini menunjukkan keberagaman data pelanggan dari sisi demografis dan pola belanja.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

Analisis korelasi antar variabel menunjukkan beberapa pola yang relevan untuk segmentasi. Terdapat hubungan negatif sebesar **-0,33** antara usia dan skor pengeluaran, mengindikasikan bahwa pelanggan yang lebih muda cenderung memiliki skor pengeluaran lebih tinggi. Korelasi mendekati nol (**-0,01**) antara pendapatan tahunan dan skor pengeluaran menunjukkan bahwa tingkat pendapatan tidak memiliki hubungan signifikan dengan pola belanja. Temuan ini memberikan wawasan awal tentang data yang digunakan serta mendukung seleksi variabel untuk proses clustering dengan metode K-Means.

3.4 Data Pre-Processing

Setelah dataset dikumpulkan, tahap berikutnya adalah analisis data dan *pre-processing*. Pada tahap ini, dilakukan beberapa langkah untuk memastikan bahwa data siap digunakan dalam proses pengelompokan.

3.4.1 Import library

Untuk melakukan pre-processing data dan mengimplementasikan algoritma K-Means, beberapa library Python yang relevan digunakan dalam penelitian ini. Proses analisis data dilakukan dengan memanfaatkan berbagai fitur dari library-library tersebut. Berikut adalah daftar pustaka Python yang digunakan beserta fungsinya:

```
# Import library yang diperlukan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn import metrics
```

Gambar 2. Import Library

a. Pandas**:** Digunakan untuk membaca dan mengelola data dalam format CSV atau Excel, serta manipulasi data berbasis DataFrame.

b. NumPy**:** Menyediakan dukungan komputasi numerik, seperti operasi matriks dan array.

c. Matplotlib**:** Digunakan untuk membuat visualisasi data seperti grafik atau diagram.

d. Seaborn**:** Membantu dalam analisis distribusi data dan visualisasi hubungan antar-variabel.

e. Sklearn**:** Sebagai library utama untuk pembelajaran mesin, digunakan untuk pemrosesan data, pelatihan model, dan evaluasi algoritma clustering.

f. KMeans dari sklearn.cluster**:** Digunakan untuk menjalankan algoritma pengelompokan K-Means pada dataset.

g. MinMaxScaler dari sklearn.preprocessing**:** Untuk normalisasi data, memastikan bahwa semua fitur berada dalam skala yang sama.

h. StandardScaler**:** Digunakan untuk standarisasi fitur dataset.

i. Metrics dari sklearn**:** Untuk evaluasi performa model clustering.

Library-library ini diimpor menggunakan platform Google Colab, yang memungkinkan eksekusi kode Python secara langsung di lingkungan berbasis cloud. Dengan pendekatan ini, pengolahan data dilakukan secara lebih terstruktur, mulai dari eksplorasi data hingga implementasi algoritma K-Means.

3.4.2 Dataset

Pada tahap awal eksplorasi data, dataset "Mall_Customers" dibaca menggunakan fungsi read_csv dari library Pandas, seperti yang terlihat pada Gambar 3. Fungsi nya adalah untuk membaca file CSV dan mengonversinya ke dalam format DataFrame, sehingga lebih mudah dikelola selama proses analisis. Setelah membaca data, data.head() digunakan untuk dapat mengetahui beberaapa

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

baris pertama dari dataset. Hal ini bertujuan untuk memahami struktur data secara cepat tanpa perlu memeriksa keseluruhan dataset. Dataset ini memiliki informasi terkait pelanggan, seperti ID, jenis kelamin, usia, pendapatan tahunan, dan skor pengeluaran mereka.

```
data = pd.read_csv("/content/Mall_Customers.csv",encoding='unicode escape')
data
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

Gambar 3. Cek Isi Data

Langkah eksplorasi data dilanjutkan dengan memeriksa jumlah baris dan kolom dataset menggunakan atribut shape, seperti yang diperlihatkan pada Gambar 4. Dari hasil tersebut, diketahui bahwa dataset memiliki 200 baris dan 5 kolom. Informasi ini memastikan bahwa data telah berhasil dimuat dan memberikan gambaran awal mengenai ukuran dataset.

```
#Rows dan Coloumn
print("Rows: {}, Coloumns: {}". format(data.shape[0], data.shape[1]))
```

```
Rows: 200, Coloumns: 5
```

Gambar 4. Cek Baris dan Kolom Data

Selanjutnya, tipe data setiap kolom diperiksa menggunakan atribut dtypes, seperti yang terlihat pada Gambar 5. Kolom "CustomerID", "Age", "Annual Income (k$)", dan "Spending Score (1-100)" bertipe int64, sedangkan kolom "Gender" bertipe object. Pemeriksaan tipe data penting untuk memastikan kesesuaian data dengan analisis yang akan dilakukan. Jika terdapat ketidaksesuaian tipe data, maka perlu dilakukan transformasi data agar sesuai dengan kebutuhan analisis.

```
#Cek data types
data.dtypes
```

	0
CustomerID	int64
Gender	object
Age	int64
Annual Income (k$)	int64
Spending Score (1-100)	int64

dtype: object

Gambar 5. Cek Tipe Data

Sebagai tahapan eksplorasi lebih lanjut, statistik deskriptif dari dataset diperoleh menggunakan fungsi describe. Output pada Gambar 6 menunjukkan informasi statistik, seperti count, mean, std, min, kuartil (25%, 50%, 75%), dan max untuk setiap variabel numerik. Sebagai contoh, rata-rata usia pelanggan adalah 38,85 tahun, dengan rentang usia dari 18 hingga 70 tahun. Sedangkan untuk pendapatan tahunan, rata-rata adalah 60,56 ribu dolar dengan standar deviasi sebesar 26,26. Statistik ini memberikan wawasan awal mengenai distribusi data serta membantu dalam mengidentifikasi potensi anomali atau langkah pre-processing tambahan yang mungkin dibutuhkan.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

```
data.describe()
```

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Gambar 6. Deskriptif Data

3.4.3 Duplikat Data

Untuk dapat melakukan pemeriksaan adanya Duplikat data pada dataset, menggunakan fungsi duplicated().sum(). Fungsi ini dapat menghitung jumlah baris yang memiliki nilai duplikat di seluruh kolom dataset. Data duplikat mengacu pada adanya dua atau lebih baris pada dataset memiliki isi data yang sepenuhnya identik.

Namun, dalam kasus ini, setelah dilakukan pemeriksaan dengan duplicated().sum(), hasilnya adalah 0 seperti terdapat pada Gambar 7, yang berarti tidak terdapat data duplikat dalam dataset.

```
#Cek duplikat
data.duplicated().sum()
```

```
0
```

Gambar 7. Cek Duplikat Data

3.4.4 Penanganan Data Hilang

Tahapan awal dalam mengksplorasi suatu data adalah dengan memeriksa keberadaan *missing value* pada dataset. Data yang hilang merupakan nilai yang tidak terisi atau tidak tersedia dalam suatu data. Berdasarkan Gambar 8, dapat dilihat bahwa tidak ada kolom pada dataset yang memiliki *missing value*. Data yang hilang dapat memengaruhi analisis data sehingga mengurangi akurasi hasil. Untuk menangani kondisi tersebut, biasanya dilakukan penghapusan baris atau kolom dengan nilai yang hilang, atau mengganti nilai yang hilang dengan estimasi tertentu.

isnull().sum() digunakan untuk mengidentifikasi jumlah nilai kosong dalam dataset dengan menjumlahkan nilai-nilai null di setiap kolom. Pada penelitian ini, karena dataset tidak memiliki *missing value*, tidak diperlukan proses penanganan tambahan untuk data hilang seperti penghapusan atau imputasi.

```
#Cek Missing Value
data.isnull().sum()
```

	0
CustomerID	0
Gender	0
Age	0
Annual Income (k$)	0
Spending Score (1-100)	0

dtype: int64

Gambar 8. Penanganan *missing value*

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

3.4.5 Transformasi Data dan Menambahkan Kolom Baru

Tahap data transformation, dilakukan untuk dapat memanipulasi data untuk mengubah format variabel sehingga dataset siap dipakai dalam proses analisis dan pemodelan. Pada penelitian ini, transformasi dilakukan dalam kolom *Gender*. Kolom ini awalnya berbentuk data kategorikal dengan nilai *Male* dan *Female*. Untuk mempermudah proses analisis berbasis numerik atau pemodelan statistik, kolom ini diubah menjadi data numerik, di mana *Male* direpresentasikan dengan angka 0 dan *Female* dengan angka 1. Transformasi seperti ini penting untuk memastikan setiap kolom dalam dataset memiliki format data yang sesuai dengan kebutuhan analisis.

```
# Mengubah kolom Gender dari kategorikal menjadi numerik
data['Gender'] = data['Gender'].map({'Male': 0, 'Female': 1})

# Menampilkan hasil transformasi kolom Gender
print(data.head())
```

```
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0           1     NaN   19                  15                      39
1           2     NaN   21                  15                      81
2           3     NaN   20                  16                       6
3           4     NaN   23                  16                      77
4           5     NaN   31                  17                      40
```

Gambar 9. Transformasi Data

Selain transformasi, dilakukan pula penambahan kolom baru untuk memperkaya informasi dalam dataset. Dalam penelitian ini, kolom baru yang ditambahkan adalah kolom *Efisiensi*. Kolom ini dihitung berdasarkan rasio antara *Spending Score (1-100)* dan *Annual Income (k$)*.

Penambahan kolom ini bertujuan untuk mengevaluasi seberapa efisien pengeluaran pelanggan terhadap pendapatan tahunan mereka. Dengan adanya kolom ini, kita dapat dengan mudah mengidentifikasi pelanggan yang memiliki pengeluaran tinggi relatif terhadap pendapatannya, yang dapat digunakan untuk analisis segmentasi pelanggan atau pengambilan keputusan bisnis.

```
# Menambahkan kolom Efisiensi
data['Efisiensi'] = data['Spending Score (1-100)'] / data['Annual Income (k$)']

# Menampilkan hasil penambahan kolom baru
print(data[['Spending Score (1-100)', 'Annual Income (k$)', 'Efisiensi']].head())
```

```
   Spending Score (1-100)  Annual Income (k$)  Efisiensi
0                      39                  15   2.600000
1                      81                  15   5.400000
2                       6                  16   0.375000
3                      77                  16   4.812500
4                      40                  17   2.352941
```

Gambar 10. Tambahkan Kolom *Efisiensi*

3.4.6 Deteksi dan Penanganan Outlier

Outlier adalah data yang mempunyai nilai signifikan beda jika dibandingkan dengan data lainnya. Dalam analisis ini, Outlier diidentifikasi menggunakan metode Interquartile Range (IQR), yang merupakan jarak antara kuartil atas (Q3) dan kuartil bawah (Q1). Nilai IQR digunakan untuk menentukan batas bawah dan batas atas data.

Proses deteksi dimulai dengan menghitung Q1 dan Q3 untuk kolom numerik Annual Income (k$). Kemudian, batas bawah ditentukan dengan formula Q1−1.5×IQRQ1 - 1.5 \times \text{IQR}Q1−1.5×IQR, dan batas atas dengan Q3+1.5×IQRQ3 + 1.5 \times \text{IQR}Q3+1.5×IQR. Data yang berada di luar batas tersebut dianggap sebagai outlier.

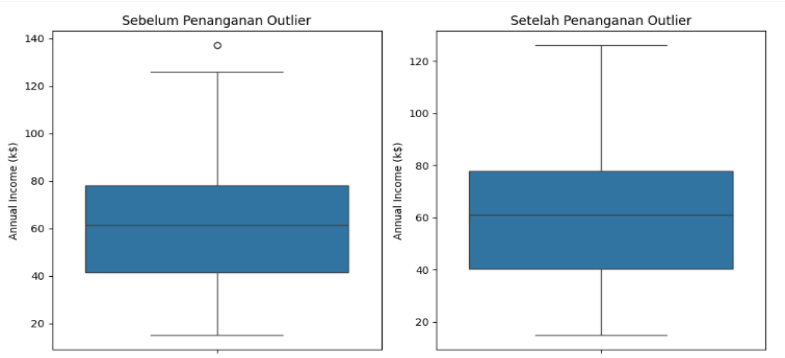
JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

```
# Deteksi dan Penanganan Outlier
Q1 = data['Annual Income (k$)'].quantile(0.25)
Q3 = data['Annual Income (k$)'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter data untuk menghapus outlier
data_cleaned = data[(data['Annual Income (k$)'] >= lower_bound) &
                    (data['Annual Income (k$)'] <= upper_bound)]
```

Gambar 11. Penanganan Outlier

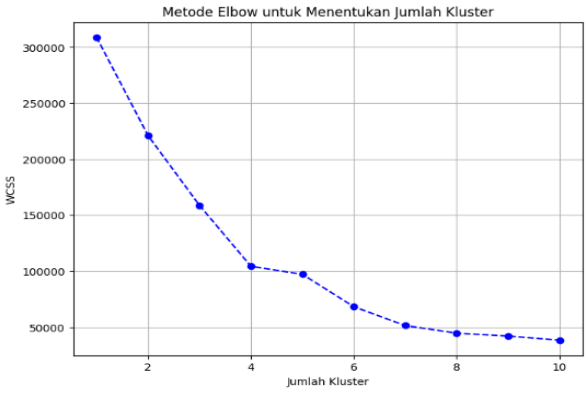
Untuk menangani outlier, data yang berada di luar rentang ini dihapus, sehingga dataset yang dihasilkan lebih bersih dan siap untuk analisis lebih lanjut. Langkah terakhir adalah memvisualisasikan data dengan diagram boxplot untuk menampilkan distribusi data sebelum dan sesudah penanganan outlier. Pada Gambar 12 ditunjukkan boxplot yang memperlihatkan adanya outlier pada data sebelum pembersihan. Dan disebelahnya menunjukkan hasil boxplot setelah data outlier dihapus, yang terlihat lebih terfokus pada rentang data normal.



Gambar 12. Diagram Boxplot

3.5 Implementasi K-Means Clustering

Setelah preprocessing dan menganalisis data, langkah berikutnya adalah mengimplementasikan algoritma K-Means untuk melakukan segmentasi pelanggan. Sebelum proses clustering dilakukan, langkah awal yang penting adalah menentukan jumlah kluster optimal. Penentuan jumlah kluster ini dilakukan menggunakan metode Elbow Method**.** Metode Elbow bertujuan untuk menemukan nilai kkk yang paling sesuai berdasarkan pengurangan nilai WCSS**.** WCSS mengukur banyaknya kuadrat jarak antara data dengan pusat klusternya. Ketika jumlah kluster bertambah, WCSS akan semakin kecil, tetapi setelah titik tertentu, penurunan WCSS tidak signifikan lagi. Titik ini disebut elbow point dan dipilih sebagai jumlah kluster optimal.



Gambar 14. Grafik Metode Elbow

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

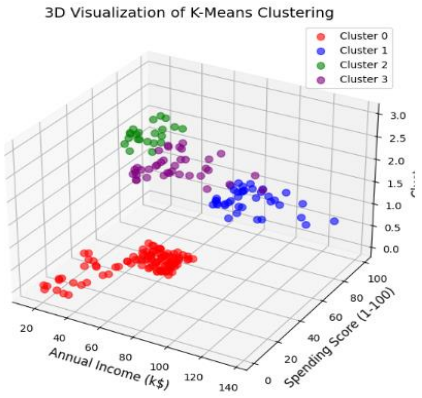
Pada tabel 3 adalah nilai WCSS dari pada cluster 1 sampai dengan 10.

Tabel 3. Nilai WCSS setiap kluster

Kluster	WCSS
1	308812.7800000001
2	221087.1962719298
3	158744.97108013942
4	104366.151455562
5	97211.84353980474
6	68275.94428646985
7	51448.36126259325
8	44640.028048530425
9	42081.855308685335
10	38378.73890793209

3.6 Hasil Clustering

Jumlah klaster yang sudah diketahui adalah 4, sehingga metode K-Means digunakan untuk membuat empat kelompok berbeda. Hasil klasterisasi yang dihasilkan divisualisasikan melalui pemetaan. Pada Gambar 15, terlihat penerapan metode K-Means yang menghasilkan 4 klaster. Visualisasi menampilkan titik-titik data yang sudah dikelompokkan ke dalam klaster dengan simbol yang berbeda-beda untuk setiap kelompok.



Pengelompokan data transaksi dalam visualisasi di atas didasarkan pada tiga variabel, yaitu Annual Income (Pendapatan Tahunan)**,** Spending Score (Skor Pengeluaran), dan dimensi tambahan sebagai indikator klaster. Berdasarkan hasil klasterisasi yang terbentuk, karakter dari setiap klaster dapat dirangkum sebagai berikut:

a. Clusters 0

Memiliki karakteristik dengan nilai *mean* kuantitas produk rendah, *mean* harga satuan tinggi, dan Customer ID dengan nilai tinggi. Kluster ini menunjukkan pelanggan yang membeli dalam jumlah sedikit tetapi dengan harga produk yang lebih mahal. Strategi yang bias dilakukan yaitu memberi diskon atau penawaran eksklusif, seperti peluncuran produk baru, untuk meningkatkan minat dan pembelian pelanggan dalam kluster ini.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 9 Februari Tahun 2025
ISSN 3025-0919 (media online)
Hal 1698-1708

b. Clusters 1

Mencakup pelanggan dengan *mean* kuantitas produk yang menengah, harga satuan yang cukup rendah, dan nilai Customer ID yang tinggi. Pelanggan di kluster ini berpotensi memiliki loyalitas yang baik. Saran yang bisa diberikan yaitu penawaran paket khusus dalam pembelian dalam jumlah besar, guna mendorong peningkatan pembelian.

c. Clusters 2

Terdiri dari pelanggan yang memiliki rata-rata kuantitas produk tinggi, harga satuan rendah, dan nilai Customer ID yang berada di tingkat menengah. Pelanggan ini cenderung membeli dalam jumlah besar dengan produk yang lebih ekonomis. Strategi yang disarankan adalah menawarkan program hadiah atau undian yang dapat meningkatkan kepuasan pelanggan dan mempertahankan hubungan jangka panjang.

d. Clusters 3

Pelanggan di kluster ini memiliki rata-rata kuantitas produk rendah, harga satuan sangat tinggi, dan nilai Customer ID yang menengah. Kluster ini mencerminkan pelanggan yang fokus pada produk premium dalam jumlah kecil. Diskon berkala atau promosi pada waktu tertentu dapat menarik minat pembelian ulang dari kelompok ini.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa segmentasi pelanggan dapat dilakukan secara efektif menggunakan algoritma K-Means Clustering dan *method* elbow dalam menentukan jumlah klaster optimal. Dengan menerapkan proses pre-processing yang mencakup deteksi outlier, normalisasi data, dan transformasi atribut, penelitian ini berhasil mengidentifikasi empat klaster pelanggan berdasarkan atribut demografi dan perilaku belanja. Klaster 0 mencerminkan pelanggan dengan pengeluaran tinggi tetapi kuantitas pembelian rendah. Klaster 1 meliputi pelanggan loyal dengan kuantitas pembelian menengah dan harga produk yang lebih rendah. Klaster 2 terdiri dari pelanggan dengan kuantitas pembelian tinggi dan produk ekonomis, sedangkan Klaster 3 mencerminkan pelanggan yang cenderung membeli produk premium dalam jumlah kecil. Hasil ini menegaskan bahwa kuantitas pembelian dan harga produk adalah faktor penting dalam memengaruhi perilaku pelanggan. Implementasi metode ini memberikan dasar untuk pengambilan keputusan strategis yang lebih efektif, serta dapat menjadi referensi untuk penelitian lebih lanjut dalam mengoptimalkan segmentasi pada pelanggan Mall menggunakan model dan algoritma lain yang lebih kompleks.

REFERENCES

Burhan, H., Kiat, Y., Azhar, & Rahmayanti, V. (2020). Penerapan metode K-Means dengan metode Elbow untuk segmentasi pelanggan menggunakan model RFM (Recency, Frequency & Monetary). *REPOSITOR*, *2*(7), 945–952.

Livari, R., & Ghalam, N. (2021). Customers grouping using data mining techniques in the food distribution industry: A case study. *SRPH Journal of Applied Management and Agile Organisation*. https://doi.org/10.47176/sjamao.3.1.1

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Juhari, T., Juarna, A., & Gunadarma, U. (2022). Implementation of RFM analysis model for customer segmentation using the K-Means algorithm: A case study at XYZ online bookstore. *EXPLORE*, *12*(1), 107–118. Retrieved June 2, 2023, from https://www.semanticscholar.org

Alamsyah, A., et al. (2022). Customer segmentation using the integration of the Recency, Frequency, Monetary model and the K-Means cluster algorithm. *Scientific Journal of Informatics*, *9*(2), 189–196. https://doi.org/10.15294/sji.v9i2.39437