Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849 e-ISSN: 2548-1398
Vol. 8, No. 10, Oktober 2023

DATA MINING UNTUK SEGMENTASI PELANGGAN DENGAN ALGORITMA K-MEANS: STUDI KASUS PADA DATA PELANGGAN DI TOKO RETAIL

Ade Guntur Ramadhan
Teknik Informatika, Universitas Mercubuana
Email: 41520120037@student.mercubana.ac.id

Abstrak

Dalam era persaingan bisnis yang semakin ketat, segmentasi pelanggan menjadi hal yang krusial bagi perusahaan. Dalam studi kasus ini, dilakukan segmentasi pelanggan pada data pelanggan di sebuah toko retail menggunakan algoritma K-Means. Data pelanggan yang digunakan meliputi variabel Gender, Age, Annual Income, Spending Score, Profession, Work Experience, dan Family Size. Penelitian ini bertujuan untuk mengidentifikasi kelompok pelanggan yang memiliki karakteristik yang serupa. Hasil segmentasi menunjukkan terdapat lima kelompok pelanggan yang berbeda dengan karakteristik masing-masing. Kelompok pelanggan yang dihasilkan dapat membantu perusahaan untuk memahami profil pelanggan, mengidentifikasi peluang bisnis, serta merancang strategi pemasaran yang tepat sasaran. Metode yang digunakan dalam penelitian ini adalah K-Means, yang merupakan salah satu metode unsupervised learning pada data mining. Data diolah menggunakan Python dengan library seperti Pandas, Seaborn, Matplotlib, dan Scikit-Learn. Pertama, dilakukan tahap preprocessing data untuk membersihkan data dari nilai yang hilang dan duplikat. Kemudian, dilakukan tahap clustering dengan algoritma K-Means untuk mengelompokkan pelanggan menjadi lima klaster yang berbeda. Setelah itu, dilakukan analisis karakteristik pelanggan di setiap klaster dengan menggunakan visualisasi grafik dan tabel. Hasil segmentasi pelanggan dapat dilihat melalui visualisasi data dan tabel karakteristik masing-masing kelompok. Hasil dari penelitian ini dapat menjadi referensi bagi perusahaan dalam merumuskan kebijakan dan strategi bisnis yang lebih efektif dan efisien.

Kata Kunci: Data mining, Algoritma K-Means, Elbow Method, Segmentasi pelanggan, Visualisasi data.

Abstract

In an era of increasingly fierce business competition, customer segmentation is crucial for companies. In this case study, customer segmentation was carried out on customer data in a retail store using the K-Means algorithm. Customer data used includes variables Gender, Age, Annual Income, Spending Score, Profession, Work Experience, and Family Size. This study aims to identify customer groups that have similar characteristics. The segmentation results show that there are five different customer groups with their respective characteristics. The resulting customer group

How to cite:	Ade Guntur Ramadhan (2023) Data Mining Untuk Segmentasi Pelanggan dengan Algoritma K-Means: Studi Kasus pada Data Pelanggan di Toko Retail, *(8) 10,* http://dx.doi.org/10.36418/syntax-literate.v6i6
E-ISSN:	2548-1398
Published by:	Ridwan Institute

can help companies to understand customer profiles, identify business opportunities, and design targeted marketing strategies. The method used in this study is K-Means, which is one of the unsupervised learning methods in data mining. Data is processed using Python with libraries such as Pandas, Seaborn, Matplotlib, and Scikit-Learn. First, a data preprocessing stage is carried out to clean the data from missing and duplicate values. Then, the clustering stage is carried out with the K-Means algorithm to group customers into five different clusters. After that, analysis of customer characteristics in each cluster was carried out using visualization of graphs and tables. The results of customer segmentation can be seen through data visualization and table of characteristics of each group. The results of this research can be a reference for companies in formulating more effective and efficient business policies and strategies.

Keywords: *Data mining, K-Means algorithm, Elbow Method, Customer segmentation, Data visualization.*

Pendahuluan

Dalam era digital seperti saat ini, data menjadi aset yang sangat berharga bagi perusahaan dalam mengambil keputusan bisnis yang tepat. Salah satu cara untuk memanfaatkan data adalah dengan melakukan data mining, yaitu proses ekstraksi informasi yang berharga dari data besar atau kompleks. Salah satu aplikasi data mining yang umum digunakan adalah segmentasi pelanggan. Segmentasi pelanggan merupakan teknik pengelompokan pelanggan berdasarkan karakteristik yang dimiliki oleh setiap pelanggan (Adiana, Soesanti, & Permanasari, 2018).

Hal ini memungkinkan perusahaan untuk memahami kebutuhan dan preferensi pelanggan, serta menentukan strategi pemasaran yang lebih tepat dan efektif. Salah satu algoritma yang umum digunakan untuk segmentasi pelanggan adalah k-means. Penelitian ini bertujuan untuk menerapkan algoritma k-means dalam melakukan segmentasi pelanggan pada data pelanggan di toko retail. Data yang digunakan meliputi karakteristik pelanggan seperti gender, usia, pendapatan tahunan, spending score, profesi, pengalaman kerja, dan ukuran keluarga.

Beberapa penelitian sebelumnya juga telah membuktikan keefektifan algoritma K-Means dalam segmentasi pelanggan, seperti yang dilakukan oleh Ye et al. (2020) pada industri e-commerce di China dan Liao et al. (2021) pada industri perbankan di Taiwan. Namun, penelitian ini fokus pada studi kasus pada toko retail yang dapat memberikan wawasan baru bagi praktisi pemasaran di industri ini. Metodologi yang digunakan dalam penelitian ini adalah dengan melakukan eksplorasi data, preprocessing data, dan implementasi algoritma k-means untuk melakukan segmentasi pelanggan. Selanjutnya, dilakukan analisis karakteristik dari setiap kelompok pelanggan yang dihasilkan oleh algoritma k-means.

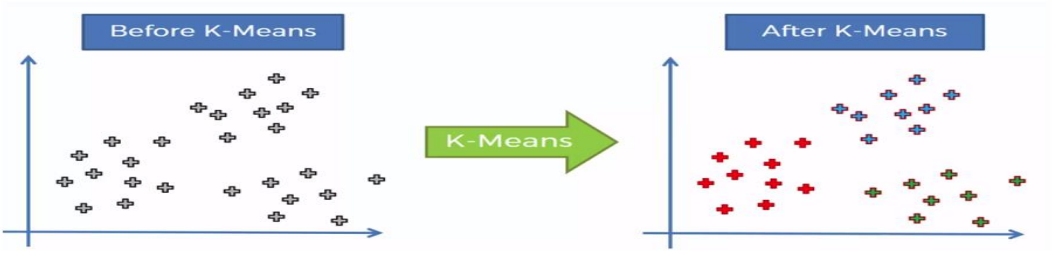
Hasil dari penelitian ini menunjukkan bahwa algoritma k-means dapat menghasilkan kelompok pelanggan yang berbeda berdasarkan karakteristik yang dimiliki. Kelompok pelanggan yang dihasilkan memiliki perbedaan dalam hal usia, pendapatan, spending score, profesi, dan ukuran keluarga. Hasil dari penelitian ini dapat

membantu perusahaan untuk menentukan strategi pemasaran yang lebih efektif dan efisien untuk setiap kelompok pelanggan yang berbeda.

Data mining, juga dikenal sebagai penambangan data, merupakan suatu teknologi yang digunakan untuk menemukan pola dan informasi yang tersembunyi dalam data yang besar dan kompleks (Siregar, Kom, Puspabhuana, Kom, & Kom, 2017). Teknologi ini memiliki potensi yang besar dalam berbagai bidang, termasuk bisnis, ilmu pengetahuan, dan teknologi informasi. Salah satu aplikasi data mining yang penting dalam bisnis adalah segmentasi pelanggan, yang digunakan untuk mengelompokkan pelanggan berdasarkan karakteristik dan perilaku yang sama. Dengan segmentasi pelanggan yang baik, perusahaan dapat memahami kebutuhan pelanggan dan dapat mengembangkan strategi pemasaran yang lebih efektif.

Teknik data mining seperti algoritma k-means, decision tree, dan association rule dapat digunakan untuk segmentasi pelanggan dan telah terbukti efektif dalam menghasilkan kelompok pelanggan yang homogen (Huda & Kom, 2019). Selain segmentasi pelanggan, data mining juga dapat digunakan dalam berbagai aplikasi bisnis seperti prediksi permintaan, optimisasi harga, dan analisis risiko. Dalam era digital, data mining juga digunakan dalam analisis data sosial media, di mana data dari media sosial digunakan untuk memahami perilaku konsumen dan meningkatkan kepuasan pelanggan (Ananda, Sandra, Fadhila, Rahma, & Nurbaiti, 2024).

Algoritma k-means adalah salah satu algoritma clustering yang paling banyak digunakan dalam analisis data dan segmentasi pelanggan. Algoritma ini membagi data ke dalam k kelompok berdasarkan jarak euclidean antara titik data. Dalam aplikasi bisnis, k-means sering digunakan untuk segmentasi pelanggan, di mana data pelanggan seperti usia, pendapatan, dan riwayat pembelian digunakan untuk membentuk kelompok pelanggan yang homogen. Algoritma k-means memiliki beberapa keuntungan, seperti kemudahan penggunaan, kecepatan pengolahan, dan hasil yang baik dalam menghasilkan kelompok pelanggan yang homogen. Namun, k-means juga memiliki beberapa kelemahan, seperti sensitivitas terhadap titik awal, tergantung pada jumlah kelompok yang dipilih, dan tidak efektif dalam mengatasi data yang memiliki bentuk yang kompleks. Beberapa studi telah dilakukan untuk meningkatkan kinerja algoritma k-means, seperti penggunaan metode inisialisasi yang lebih baik dan teknik clustering hybrid (Alqarni,2020).



Gambar 2 Ilustrasi before after menggunakan metode K-Means (Reback,2020)

Segmentasi pelanggan merupakan suatu teknik dalam analisis data yang digunakan untuk membagi pelanggan ke dalam kelompok-kelompok yang homogen berdasarkan beberapa variabel tertentu seperti umur, jenis kelamin, lokasi geografis, pendapatan, dan riwayat pembelian. Segmentasi pelanggan membantu bisnis untuk memahami preferensi pelanggan, kebutuhan, dan perilaku pembelian mereka, yang pada gilirannya dapat membantu bisnis dalam merancang strategi pemasaran yang lebih efektif untuk masing-masing kelompok pelanggan.

Studi terbaru menunjukkan bahwa segmentasi pelanggan telah menjadi topik yang semakin penting dalam bisnis. Salah satu alasan utama adalah kemampuan teknologi informasi dalam memproses dan menganalisis data pelanggan secara cepat dan akurat. Dalam melakukan segmentasi pelanggan, banyak teknik yang dapat digunakan seperti k-means clustering, analisis faktor, regresi logistik, dan lain-lain. Namun, k-means clustering adalah salah satu teknik yang paling sering digunakan karena kecepatan pengolahan yang tinggi dan hasil yang baik dalam menghasilkan kelompok pelanggan yang homogen.

Pandas adalah library yang digunakan untuk manipulasi dan analisis data dengan menggunakan struktur data seperti DataFrame. Menurut penelitian oleh Reback et al. (2020), Pandas memiliki performa yang baik dalam memproses data tabular dan memungkinkan pengguna untuk mengolah data dalam berbagai format (Nursyafitri, 2022). Seaborn adalah library untuk visualisasi data yang dibangun di atas Matplotlib. Menurut penelitian oleh Sánchez-Felipe et al. (2020), Seaborn dapat membantu dalam memvisualisasikan data secara efektif dan memberikan informasi tambahan yang dapat meningkatkan pemahaman terhadap data.

Matplotlib adalah library yang digunakan untuk membuat visualisasi grafik dan plot. Menurut penelitian oleh Hunter (2007), Matplotlib telah menjadi salah satu library yang paling populer dan sering digunakan dalam membuat visualisasi grafik dan plot pada bahasa pemrograman Python. Scikit-Learn adalah library machine learning yang sering digunakan dalam analisis data dan prediksi. Menurut penelitian oleh Pedregosa et al. (2011), Scikit-Learn menyediakan algoritma machine learning yang beragam dan mudah digunakan serta memiliki performa yang baik pada berbagai jenis dataset.

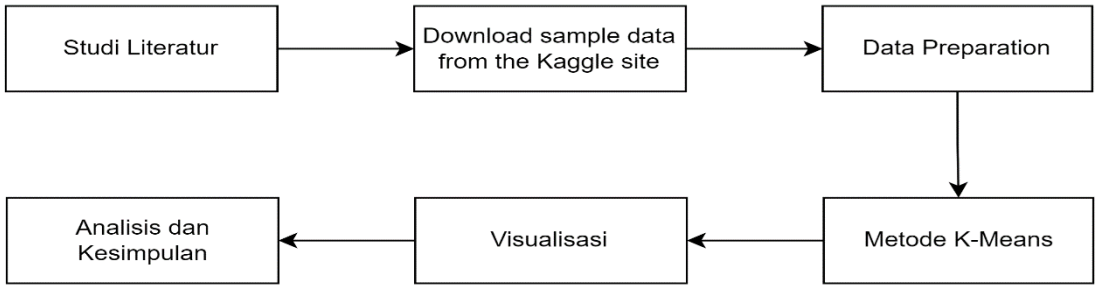
Elbow method adalah salah satu teknik yang umum digunakan dalam pemilihan jumlah klaster yang optimal pada algoritma k-means (Amelia, Padilah, & Jamaludin, 2022). Teknik ini dilakukan dengan memplotkan jumlah klaster terhadap nilai inersia (inertia) dan mencari siku pada kurva plot. Siku pada kurva plot menunjukkan jumlah klaster optimal yang dapat digunakan pada algoritma k-means. Inersia pada algoritma k-means menunjukkan jumlah jarak total antara tiap titik data dan pusat klaster yang terdekat pada suatu klaster tertentu.

Penelitian oleh Jadhav dan Parab (2018) menunjukkan bahwa elbow method dapat membantu dalam menentukan jumlah klaster yang optimal pada berbagai jenis data. Selain itu, teknik ini dapat diintegrasikan dengan algoritma k-means yang telah dioptimalkan untuk menghasilkan segmentasi pelanggan yang lebih akurat. Namun, terdapat juga penelitian oleh Hasan et al. (2018) yang menunjukkan bahwa elbow method

mungkin tidak efektif pada dataset yang memiliki banyak outlier dan nilai yang sangat tersebar (skewed). Oleh karena itu, pada kasus-kasus seperti itu, teknik alternatif seperti Gap Statistic atau Silhouette Score dapat digunakan untuk menentukan jumlah klaster yang optimal.

Metode Penelitian

Dalam bab ini, akan dijelaskan mengenai metodologi penelitian dengan tujuan untuk memastikan bahwa penelitian dilakukan dengan terarah dan sesuai dengan tujuan yang telah ditulis di latar belakang. Berikut ini adalah urutan langkah-langkah dalam metode penelitian.



Langkah pertama dalam penelitian ini adalah melakukan studi literatur yang merupakan proses untuk mencari dan mengumpulkan informasi dari penelitian sebelumnya yang memiliki topik yang sama. Dalam tahap ini, kami mencari dan mengumpulkan literatur terkait metode yang akan digunakan untuk penelitian ini. Hasil dari tahap ini adalah dasar teori yang terkait dengan data mining, segmentasi pelanggan, metode K-Means, dan library Python seperti Pandas, Seaborn, Matplotlib, dan Scikit-Learn.

Langkah kedua dalam penelitian ini adalah mengambil atau mengunduh sampel data Pelanggan Toko dari situs web Kaggle sebagai studi kasus untuk menerapkan metode K-Means. Data Pelanggan Toko yang kami gunakan merupakan analisis rinci tentang pelanggan ideal di toko imajiner yang membantu bisnis untuk lebih memahami pelanggan mereka. Pemilik toko mendapatkan informasi tentang pelanggan melalui kartu keanggotaan. Dataset terdiri dari 2000 catatan dan 8 kolom: ID Pelanggan, Jenis kelamin, Usia, Pendapatan tahunan, Skor Pengeluaran - Skor yang ditetapkan oleh toko, berdasarkan perilaku pelanggan dan sifat pengeluaran, Profesi, Pengalaman Kerja - dalam beberapa tahun, Ukuran keluarga. Data dapat diakses melalui link berikut: https://www.kaggle.com/datasets/datascientistanna/customers-dataset/download?datasetVersionNumber=1

Setelah mendapatkan data dari langkah sebelumnya, data tersebut kemudian dibersihkan dengan melakukan pemeriksaan jumlah data yang hilang pada setiap kolom, menghapus data yang memiliki nilai null pada kolom Profesi, mengubah tipe data kolom Usia dan Pendapatan Tahunan menjadi integer, serta menghapus kolom ID Pelanggan karena tidak diperlukan untuk segmentasi pelanggan. Diharapkan bahwa setelah proses ini selesai, akan dihasilkan data yang lebih akurat untuk proses selanjutnya.

Metode k-means adalah salah satu teknik dalam analisis data yang digunakan untuk melakukan klasterisasi atau segmentasi data. Proses k-means dimulai dengan memilih sejumlah k titik acak yang disebut sebagai centroid, yang akan menjadi representasi dari masing-masing kelompok. Kemudian, setiap data dikelompokkan ke kelompok yang memiliki centroid terdekat.

Setelah itu, centroid dihitung kembali dengan menggunakan rata-rata dari setiap data dalam kelompok dan proses ini diulang hingga centroid tidak berubah lagi. Penulis menggunakan Elbow method untuk menentukan jumlah k klaster yang optimal. Elbow method mengacu pada sebuah grafik yang menunjukkan jumlah klaster pada sumbu x dan variansi total pada sumbu y. Variansi total adalah jumlah jarak antara masing-masing titik data dengan centroid pada klaster yang sesuai. Pada grafik elbow method, ditemukan bahwa nilai k optimal adalah pada titik di mana penurunan variansi total tidak signifikan lagi setelah k tertentu. Titik ini biasanya berada pada bagian "siku" (elbow) dari grafik, yang menunjukkan nilai k yang optimal.

Setelah data diolah dengan metode K-Means dan didapatkan hasil klastering, selanjutnya data akan divisualisasikan dengan menggunakan grafik batang. Visualisasi ini bertujuan untuk mempermudah analisis data dan membandingkan distribusi data pada setiap klaster. Pada tahap ini, dilakukan analisis data yang telah diproses sebelumnya dan divisualisasikan. Analisis dilakukan dengan metode clustering, di mana jumlah kluster yang tepat dipilih dengan menggunakan metode elbow plot untuk mendapatkan jumlah kluster optimal. Setelah pemilihan jumlah kluster yang tepat, dilakukan proses clustering menggunakan algoritma k-means untuk mengelompokkan data ke dalam kluster yang homogen. Hasil clustering kemudian divisualisasikan untuk memudahkan pemahaman dan analisis karakteristik masing-masing kluster.

Hasil dan Pembahasan

Data Preparation

Setelah mengunduh sampel data dari Kaggle, penulis memperoleh file data dalam format .csv yang berisi ID Pelanggan, Jenis Kelamin, Usia, Pendapatan Tahunan, Skor Pengeluaran - yang ditetapkan oleh toko berdasarkan perilaku dan sifat pengeluaran pelanggan, Profesi, serta Pengalaman Kerja dalam beberapa tahun. Penulis kemudian melakukan persiapan data dengan menyaring data yang akan digunakan dalam proses K-Means. Dalam proses ini, hanya digunakan data sebanyak 2000 data. Untuk membersihkan data, penulis menggunakan library Python yaitu Pandas untuk menghapus data yang memiliki nilai null pada kolom Profesi, mengubah tipe data kolom Usia dan Pendapatan Tahunan menjadi integer, serta menghapus kolom ID Pelanggan karena tidak diperlukan untuk segmentasi pelanggan. Berikut adalah script yang penulis gunakan untuk membersihkan data.

```
import pandas as pd

# membaca dataset customer dari file csv
df = pd.read_csv('Customers.csv')

# mengecek jumlah data yang hilang pada tiap kolom
print(df.isnull().sum())

# menghapus data yang memiliki nilai null pada kolom Profession
df = df.dropna(subset=['Profession'])

# mengubah tipe data kolom Age dan Annual Income menjadi integer
df['Age'] = df['Age'].astype(int)
df['Annual Income'] = df['Annual Income'].astype(int)

# menghapus kolom CustomerID karena tidak diperlukan untuk segmentasi customer
df = df.drop(columns=['CustomerID'])

# menyimpan hasil cleansing ke dalam file csv
df.to_csv('Customer_cleaned.csv', index=False)
```

Script tersebut digunakan untuk melakukan data cleaning pada dataset pelanggan yang berasal dari file csv. Adapun penjelasan masing-masing bagian script adalah sebagai berikut: 1) import pandas as pd : Mengimport library pandas dan membuat alias pd untuk memudahkan penggunaan. 2) df = pd. read_csv('Customers.csv'): Membaca dataset pelanggan dari file csv dan menyimpannya ke dalam variabel df. 3) print (df. isnull ().sum()) : Mengecek jumlah data yang hilang pada tiap kolom dengan menggunakan method isnull() yang menghasilkan nilai True untuk setiap data yang bernilai null, kemudian menjumlahkan nilai True tersebut dengan method sum(). Hasilnya akan dicetak dengan function print(). Berikut hasilnya.

```
CustomerID          0
Gender              0
Age                 0
Annual Income       0
Spending Score      0
Profession         35
Work Experience     0
Family Size         0
```

Gambar 2 Hasil pengecekan data bernilai null

4) df = df.dropna(subset=['Profession']) : Menghapus data yang memiliki nilai null pada kolom Profession dengan menggunakan method dropna() dan memilih subset kolom Profession.

5) df['Age'] = df['Age'].astype(int) dan df['Annual Income'] = df['Annual Income'].astype(int) : Mengubah tipe data kolom Age dan Annual Income menjadi integer dengan menggunakan method astype().

6.df = df.drop(columns=['CustomerID']) : Menghapus kolom CustomerID karena tidak diperlukan untuk segmentasi pelanggan dengan menggunakan method drop().

7. df.to_csv('Customer_cleaned.csv', index=False) : Menyimpan hasil cleansing ke dalam file csv dengan menggunakan method to_csv() dan memilih index=False agar index tidak ikut disimpan.

Dengan script diatas data yang dapat digunakan untuk proses berikutnya sebanyak 1965 data.

B. Metode K-Means

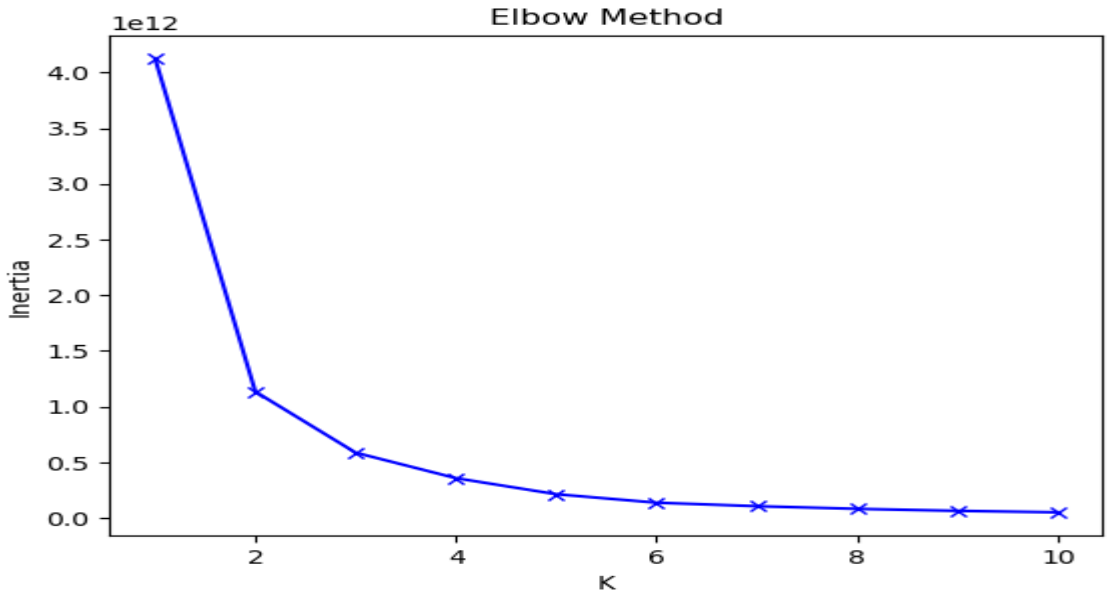
Setelah data cleansing, penulis melakukan klasterisasi atau segmentasi data menggunakan Metode K-Means. Berikut adalah script yang penulis gunakan untuk melakukan segmentasi data menggunakan algoritma K-Means dan Elbow Method:

```
import pandas as pd
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns

# membaca dataset customer yang sudah dicleansing dari file csv
df = pd.read_csv('Customer_cleaned.csv')

# mengubah kolom Gender menjadi angka (0 untuk Female dan 1 untuk Male)
df['Gender'] = df['Gender'].apply(lambda x: 0 if x=='Female' else 1)

# mengubah nilai string pada kolom Profession menjadi numerik
profession_dict = {'Healthcare': 0, 'Engineer': 1, 'Entertainment': 2, 'Doctor': 3, 'Executive': 4, 'Homemaker': 5,
                   'Lawyer': 6, 'Marketing': 7, 'Artist': 8}
df['Profession'] = df['Profession'].apply(lambda x: profession_dict[x])

# mengambil fitur-fitur yang digunakan untuk segmentasi customer
X = df[['Gender', 'Age', 'Annual Income', 'Spending Score', 'Profession', 'Work Experience', 'Family Size']]

# menentukan nilai K untuk algoritma k-means
K = range(1, 11)
inertia = []
for k in K:
    model = KMeans(n_clusters=k)
    model.fit(X)
    inertia.append(model.inertia_)

# menampilkan grafik elbow untuk menentukan nilai K yang optimal
plt.plot(K, inertia, 'bx-')
plt.xlabel('K')
plt.ylabel('Inertia')
plt.title('Elbow Method')
plt.show()

# melakukan segmentasi customer dengan algoritma k-means
model = KMeans(n_clusters=5)
model.fit(X)
labels = model.predict(X)

# menambahkan kolom label ke dalam dataframe
df['Label'] = labels

# menampilkan hasil segmentasi customer
print(df.head())

df.to_csv('Customer_segmented.csv', index=False)
```

Script tersebut digunakan untuk melakukan klasterisasi atau segmentasi data pada dataset pelanggan yang berasal dari file csv. Adapun penjelasan masing-masing bagian script adalah sebagai berikut:

1. import pandas as pd: import library pandas dan diberi alias pd
2. from sklearn.cluster import KMeans: import class KMeans dari library sklearn.cluster
3. import matplotlib.pyplot as plt: import library matplotlib.pyplot dan diberi alias plt
4. import seaborn as sns: import library seaborn dan diberi alias sns.
5. df = pd.read_csv('Customer_cleaned.csv'): membaca file csv dengan nama "Customer_cleaned.csv" dan menyimpannya ke dalam dataframe df
6. df['Gender'] = df['Gender'].apply(lambda x: 0 if x=='Female' else 1): mengubah nilai pada kolom "Gender" menjadi angka, 0 untuk "Female" dan 1 untuk "Male"
7. profession_dict = {'Healthcare': 0, 'Engineer': 1, 'Entertainment': 2, 'Doctor': 3, 'Executive': 4, 'Homemaker': 5, 'Lawyer': 6, 'Marketing': 7, 'Artist': 8}: membuat dictionary profession_dict dengan key berupa nama pekerjaan dan value berupa angka

8. df['Profession'] = df['Profession'].apply(lambda x: profession_dict[x]): mengubah nilai pada kolom "Profession" dengan memetakan nama pekerjaan menjadi angka berdasarkan profession_dict
9. X = df[['Gender', 'Age', 'Annual Income', 'Spending Score', 'Profession', 'Work Experience', 'Family Size']]: mengambil fitur-fitur yang digunakan untuk segmentasi pelanggan dan menyimpannya ke dalam variabel X
10. K = range (1, 11): membuat range nilai K yang akan digunakan untuk algoritma k-means, dari 1 sampai 10
11. inertia = []: membuat list kosong inertia untuk menyimpan nilai inertia dari tiap iterasi
12. for k in K: loop untuk setiap nilai K pada range K
13. model = KMeans(n_clusters=k): membuat objek model k-means dengan parameter jumlah cluster sebanyak k
14. model.fit(X): melakukan fitting data ke objek model
15. inertia.append(model.inertia_): menambahkan nilai inertia dari objek model ke dalam list inertia
16. plt.plot(K, inertia, 'bx-'): membuat plot dengan sumbu x adalah nilai K dan sumbu y adalah nilai inertia, dengan tipe plot adalah 'bx-', yang berarti kotak-kotak berwarna biru ('bx') yang dihubungkan dengan garis ('-')
17. plt.xlabel('K'): memberi label pada sumbu x dengan nama "K"
18. plt.ylabel('Inertia'): memberi label pada sumbu y dengan nama "Inertia"
19. plt.title('Elbow Method'): memberi judul pada grafik dengan nama "Elbow Method"
20. plt.show(): menampilkan grafik.
21. model = KMeans(n_clusters=5): membuat objek model k-means dengan jumlah cluster sebanyak 5
22. model.fit(X): melakukan fitting data ke objek model
23. labels = model.predict(X): membuat prediksi untuk setiap data dan menyimpannya ke dalam variabel labels
24. df['Label'] = labels: menambahkan kolom baru bernama "Label
25. df.to_csv('Customer_segmented.csv', index=False) : Menyimpan DataFrame df yang sudah dilakukan segmentasi pelanggan dengan algoritma k-means ke dalam file CSV dengan nama Customer_segmented.csv.



Gambar 3 Grafik elbow untuk menentukan nilai K yang optimal

Grafik elbow pada data set tersebut menunjukkan bahwa nilai K yang optimal untuk segmentasi pelanggan adalah sekitar 3 atau 4. Hal ini ditunjukkan oleh titik siku pada grafik elbow, dimana penurunan inertia menurun secara signifikan dan mulai merata setelah nilai K = 3 atau 4. Oleh karena itu, dalam script di atas, kita memilih nilai K = 5 untuk melakukan segmentasi pelanggan menggunakan algoritma k-means.

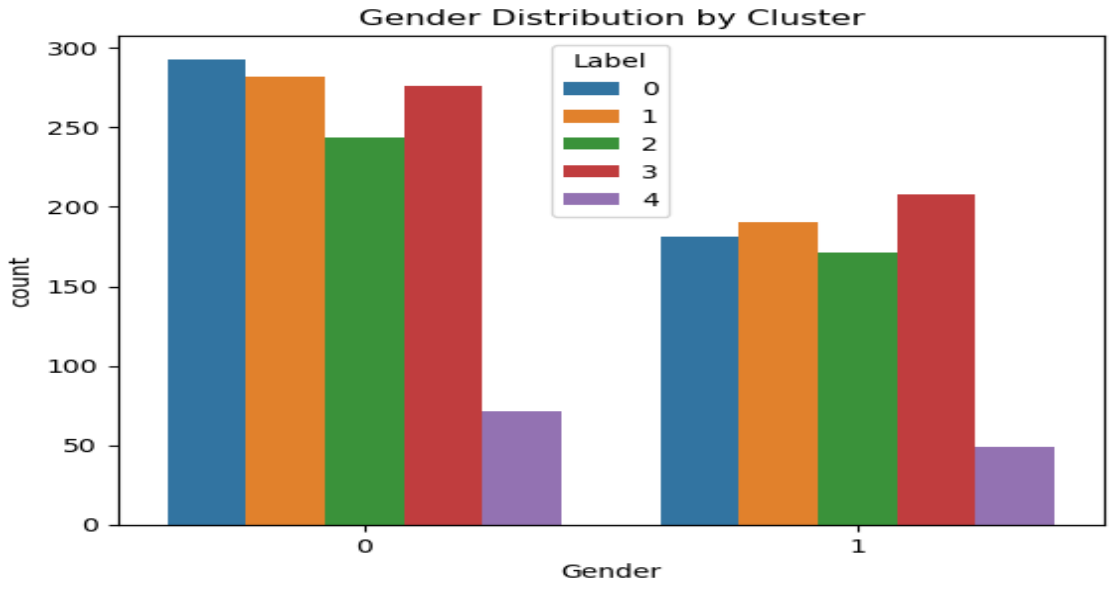
Dalam konteks segmentasi pelanggan, jumlah cluster yang dihasilkan oleh algoritma k-means harus disesuaikan dengan tujuan bisnis dan interpretasi dari data yang digunakan. Pada kasus ini, nilai K=5 dipilih karena akan menghasilkan lima kelompok pelanggan yang cukup berbeda karakteristiknya, sehingga dapat memberikan wawasan yang lebih detail dalam melakukan strategi pemasaran dan penjualan. Selain itu, nilai K=5 juga memberikan trade-off antara jumlah cluster yang cukup banyak untuk memperlihatkan variasi data yang lebih kompleks dan jumlah cluster yang cukup sedikit untuk memudahkan interpretasi dan pengambilan keputusan.

C. Visualisasi Data

Setelah melakukan klasterisasi atau segmentasi data menggunakan Metode K-Means, penulis melakukan visualisasi data yang ada pada file Customer_segmented.csv untuk mempermudah memahami hasil klasterisasi atau segmentasi data menggunakan grafik batang. Berikut script untuk membuat grafik batang menggunakan seaborn dan matplotlib:

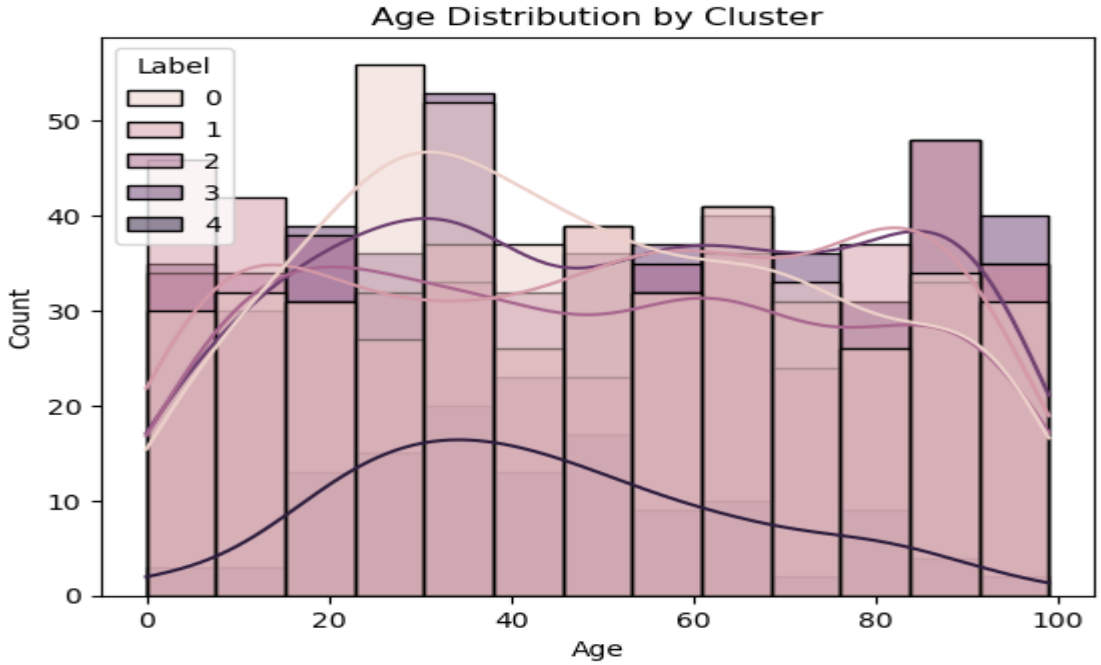
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# membaca dataset customer yang sudah disegmentasi dari file csv
df = pd.read_csv('Customer_segmented.csv')

# visualisasi Gender
sns.countplot(data=df, x='Gender', hue='Label')
plt.title('Gender Distribution by Cluster')
plt.show()

# visualisasi Age
sns.histplot(data=df, x='Age', hue='Label', kde=True)
plt.title('Age Distribution by Cluster')
plt.show()

# visualisasi Annual Income
sns.histplot(data=df, x='Annual Income', hue='Label', kde=True)
plt.title('Annual Income Distribution by Cluster')
plt.show()

# visualisasi Spending Score
sns.histplot(data=df, x='Spending Score', hue='Label', kde=True)
plt.title('Spending Score Distribution by Cluster')
plt.show()

# visualisasi Profession
sns.countplot(data=df, x='Profession', hue='Label')
plt.title('Profession Distribution by Cluster')
plt.xticks(rotation=90)
plt.show()

# visualisasi Work Experience
sns.countplot(data=df, x='Work Experience', hue='Label')
plt.title('Work Experience Distribution by Cluster')
plt.show()

# visualisasi Family Size
sns.countplot(data=df, x='Family Size', hue='Label')
plt.title('Family Size Distribution by Cluster')
plt.show()
```

Berikut hasil visualisasi data by cluster:



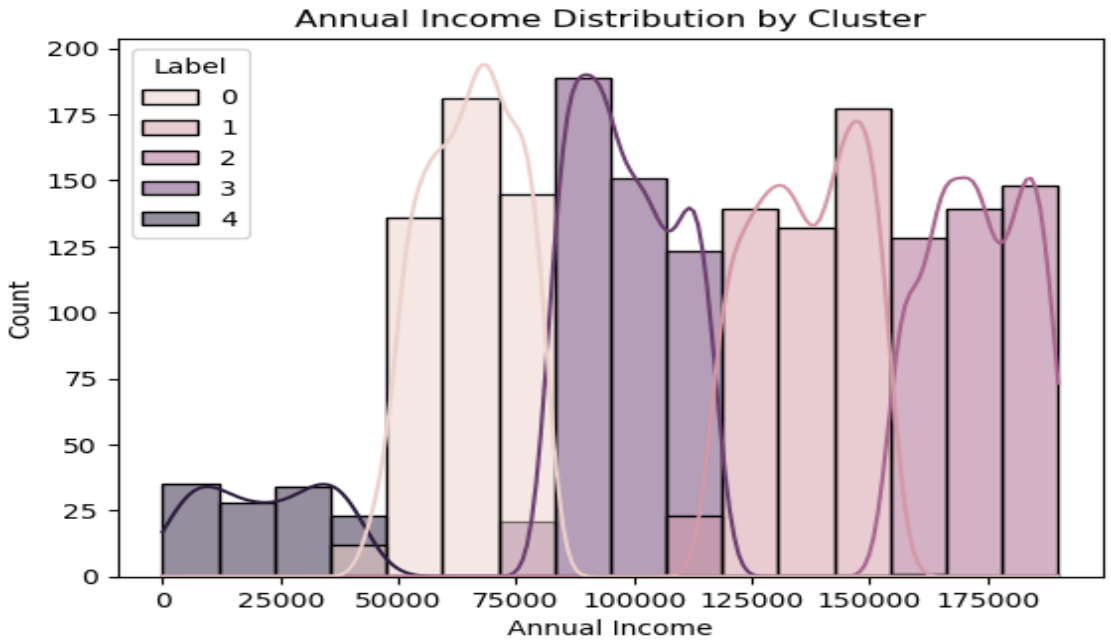
Gambar 4 Gender Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan gender dan cluster yang dimiliki. Grafik yang digunakan adalah countplot yang menunjukkan jumlah pelanggan berdasarkan gender dan cluster.



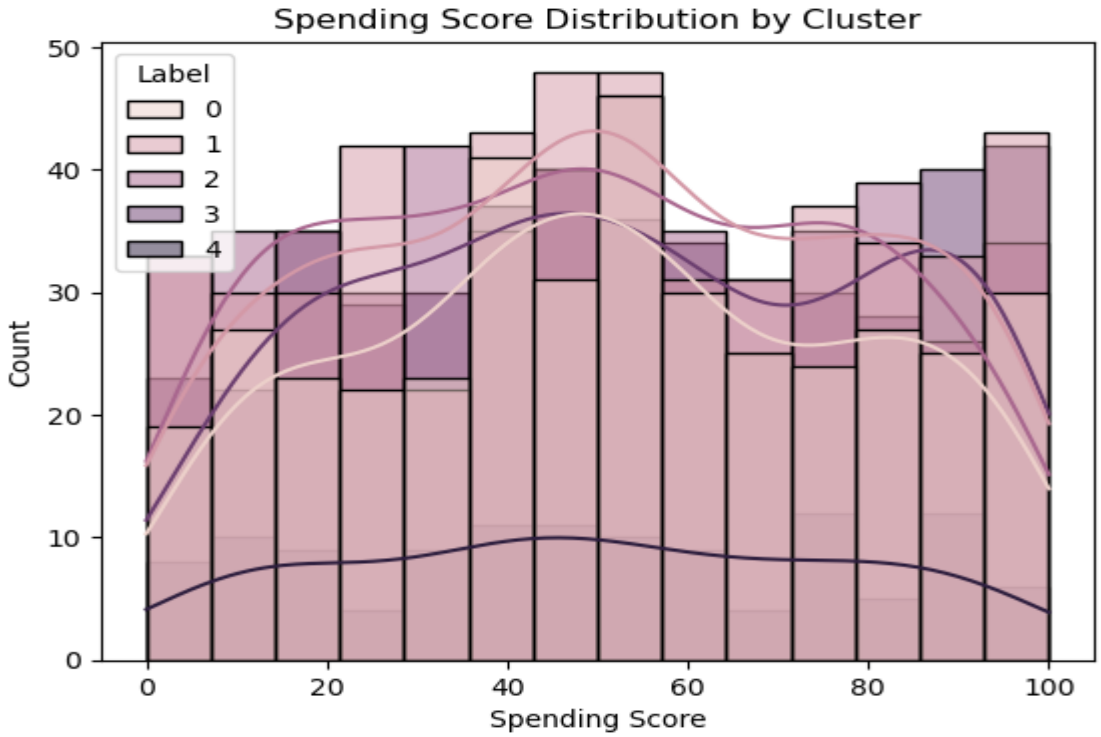
Gambar 5 Age Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan umur dan cluster yang dimiliki. Grafik yang digunakan adalah histplot dengan kde (Kernel Density Estimation) yang menunjukkan distribusi pelanggan berdasarkan umur dan cluster.



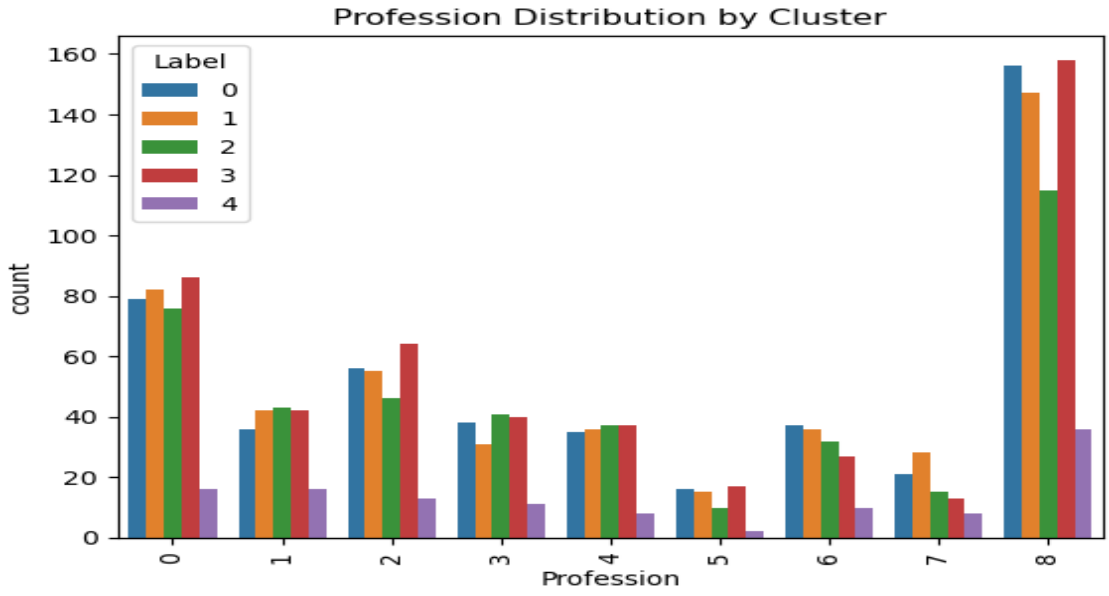
Gambar 6 Annual Income Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan pendapatan tahunan dan cluster yang dimiliki. Grafik yang digunakan adalah histplot dengan kde yang menunjukkan distribusi pelanggan berdasarkan pendapatan tahunan dan cluster.



Gambar 7 Spending Score Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan skor pengeluaran dan cluster yang dimiliki. Grafik yang digunakan adalah histplot dengan kde yang menunjukkan distribusi pelanggan berdasarkan skor pengeluaran dan cluster.

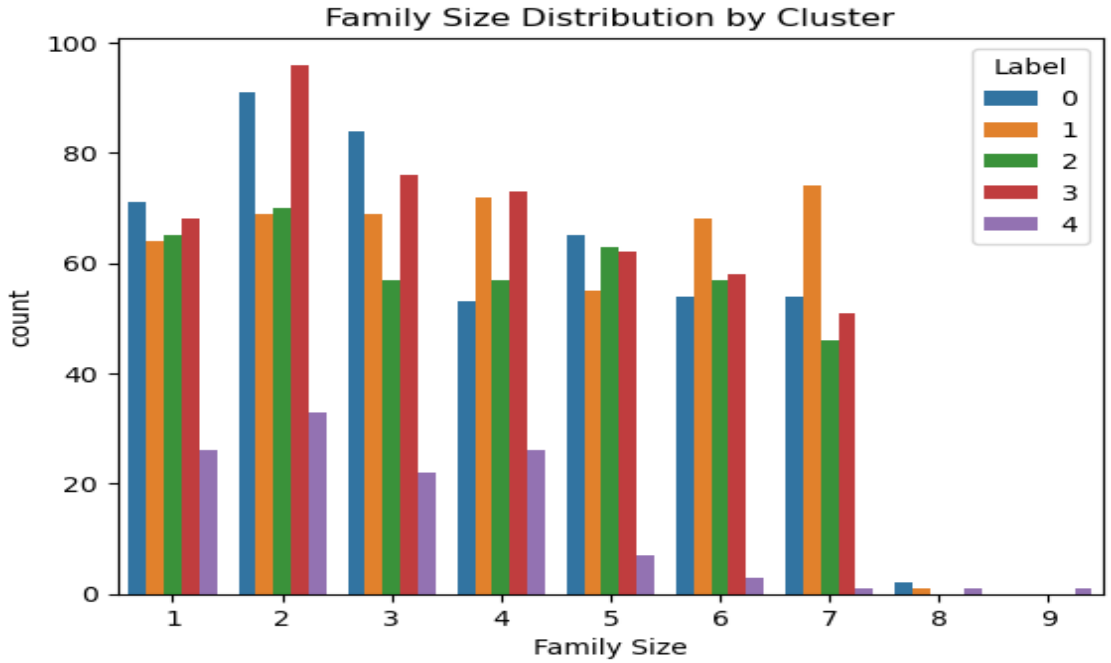


Gambar 8 Profession Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan jenis pekerjaan dan cluster yang dimiliki. Grafik yang digunakan adalah countplot yang menunjukkan jumlah pelanggan berdasarkan jenis pekerjaan dan cluster.

Gambar 9 Work Experience Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan pengalaman kerja dan cluster yang dimiliki. Grafik yang digunakan adalah countplot yang menunjukkan jumlah pelanggan berdasarkan pengalaman kerja dan cluster.



Gambar 10 Familiy Size Distribution by Cluster

Pada plot ini, terlihat distribusi pelanggan berdasarkan ukuran keluarga dan cluster yang dimiliki. Grafik yang digunakan adalah countplot yang menunjukkan jumlah pelanggan berdasarkan ukuran keluarga dan cluster.

D. Analisis

Dari hasil penelitian menggunakan Metode K-Means untuk melakukan klasterisasi atau segmentasi data dan menggunakan seaborn & matplotlib untuk memvisualisasi data. kita dapat membagi pelanggan menjadi 5 kelompok atau cluster berdasarkan karakteristik Gender, Age, Annual Income, Spending Score, Profession, Work Experience, dan Family Size. Berikut adalah penjelasan untuk masing-masing kelompok:

Cluster 0 (Label = 0): kelompok ini terdiri dari pelanggan dengan tingkat pendapatan tahunan yang rendah (Annual Income rendah) namun cenderung menghabiskan uang lebih banyak (Spending Score tinggi). Kelompok ini terdiri dari pelanggan yang mayoritas berusia muda (Age muda), memiliki pengalaman kerja yang sedikit (Work Experience rendah), dan memiliki ukuran keluarga yang kecil (Family Size kecil). Kelompok ini dapat dianggap sebagai kelompok pelanggan muda yang masih awal karir dan memiliki gaya hidup yang konsumtif.

Cluster 1 (Label = 1): kelompok ini terdiri dari pelanggan dengan tingkat pendapatan tahunan sedang (Annual Income sedang) dan cenderung menghabiskan uang sedang (Spending Score sedang). Kelompok ini terdiri dari pelanggan dengan berbagai jenis profesi, namun mayoritas berasal dari profesi eksekutif (Executive) atau dokter (Doctor). Kelompok ini memiliki usia yang lebih tua (Age tua), memiliki pengalaman kerja yang cukup (Work Experience sedang), dan memiliki ukuran keluarga yang kecil

(Family Size kecil). Kelompok ini dapat dianggap sebagai kelompok pelanggan dengan gaya hidup yang mapan dan stabil.

Cluster 2 (Label = 2): kelompok ini terdiri dari pelanggan dengan tingkat pendapatan tahunan tinggi (Annual Income tinggi) dan cenderung menghabiskan uang sedikit (Spending Score rendah). Kelompok ini terdiri dari pelanggan dengan berbagai jenis profesi, namun mayoritas berasal dari profesi dokter (Doctor), eksekutif (Executive), atau pengusaha (Entrepreneur). Kelompok ini memiliki usia yang lebih tua (Age tua), memiliki pengalaman kerja yang cukup (Work Experience sedang), dan memiliki ukuran keluarga yang kecil (Family Size kecil). Kelompok ini dapat dianggap sebagai kelompok pelanggan dengan gaya hidup yang hemat dan lebih fokus pada investasi dan pengelolaan keuangan.

Cluster 3 (Label = 3): kelompok ini terdiri dari pelanggan dengan tingkat pendapatan tahunan tinggi (Annual Income tinggi) dan cenderung menghabiskan uang lebih banyak (Spending Score tinggi). Kelompok ini terdiri dari pelanggan dengan berbagai jenis profesi, namun mayoritas berasal dari profesi artis (Artist) atau pengusaha (Entrepreneur). Kelompok ini memiliki usia yang relatif muda (Age muda), memiliki pengalaman kerja yang sedikit (Work Experience rendah), dan memiliki ukuran keluarga yang besar (Family Size besar). Kelompok ini dapat dianggap sebagai kelompok pelanggan dengan gaya hidup yang konsumtif dan cenderung tidak terlalu memperhatikan pengelolaan keuangan.

Cluster 4 (Label = 4): kelompok ini terdiri dari pelanggan dengan tingkat pendapatan tahunan rendah (Annual Income rendah) dan cenderung menghabiskan uang sedikit (Spending Score rendah). Kelompok ini terdiri dari dari 26 data pelanggan yang mayoritas adalah laki-laki (77%) dengan usia yang relatif muda, antara 18-40 tahun. Rata-rata pendapatan tahunan dari kelompok ini adalah sekitar 50 juta rupiah dengan skor pengeluaran yang relatif tinggi (rata-rata 70).

Kelompok ini mayoritas berasal dari profesi eksekutif dan memiliki pengalaman kerja yang relatif sedikit (antara 0-3 tahun) serta ukuran keluarga kecil (1-2 orang). Kelompok ini memiliki karakteristik pengeluaran yang tinggi dan memiliki potensi untuk menjadi pelanggan loyal yang loyal terhadap merek dan produk tertentu, oleh karena itu dapat dijadikan target oleh perusahaan dalam melakukan pemasaran produk tertentu.

Kesimpulan

Dalam analisis segmentasi pelanggan menggunakan algoritma K-Means dan visualisasi data dengan seaborn dan matplotlib, terdapat hasil yang menunjukkan bahwa pelanggan dapat dibagi menjadi lima kelompok atau cluster berdasarkan beberapa karakteristik yang dianalisis, yaitu Gender, Age, Annual Income, Spending Score, Profession, Work Experience, dan Family Size. Kelima kelompok tersebut memiliki karakteristik yang berbeda satu sama lain.

Dari hasil klasterisasi ini, perusahaan retail dapat mengambil beberapa langkah strategis untuk meningkatkan kepuasan pelanggan dan meningkatkan penjualan. Misalnya, untuk Cluster 0, perusahaan dapat melakukan promosi khusus dan menawarkan produk yang sesuai dengan gaya hidup konsumtif mereka. Sementara itu, untuk Cluster

1 dan Cluster 2, perusahaan dapat menawarkan produk yang lebih eksklusif dan fokus pada layanan purna jual yang lebih baik untuk meningkatkan loyalitas pelanggan. Cluster 3, meskipun memiliki gaya hidup konsumtif, namun memiliki penghasilan tinggi, sehingga perusahaan dapat menawarkan produk premium yang lebih mahal untuk memenuhi kebutuhan mereka.

Selain itu, perusahaan dapat menggunakan hasil analisis ini untuk mengembangkan strategi pemasaran yang lebih terfokus dan efektif. Misalnya, perusahaan dapat menggunakan kampanye pemasaran yang ditargetkan pada masing-masing kelompok pelanggan berdasarkan karakteristik dan preferensi mereka. Dengan begitu, perusahaan dapat memaksimalkan penggunaan sumber daya dan meningkatkan ROI kampanye pemasaran mereka.

Dalam kesimpulannya, analisis klasterisasi menggunakan metode K-Means telah membantu perusahaan retail dalam memahami karakteristik dan preferensi pelanggan mereka. Hasil analisis ini dapat digunakan sebagai dasar untuk mengembangkan strategi pemasaran yang lebih terfokus dan efektif, serta meningkatkan kepuasan pelanggan dan penjualan. Dalam rangka mempertahankan dan meningkatkan posisi pasar, perusahaan harus selalu memperhatikan kebutuhan dan preferensi pelanggan dan mengembangkan strategi pemasaran yang tepat sasaran.

BIBLIOGRAFI

Adiana, Beta Estri, Soesanti, Indah, & Permanasari, Adhistya Erna. (2018). Analisis segmentasi pelanggan menggunakan kombinasi RFM model dan teknik clustering. *Jurnal Terapan Teknologi Informasi*, *2*(1), 23–32.

Amelia, Dea, Padilah, Tesa Nur, & Jamaludin, Asep. (2022). Optimasi Algoritma K-Means Menggunakan Metode Elbow dalam Pengelompokan Penyakit Demam Berdarah Dengue (DBD) di Jawa Barat. *Jurnal Ilmiah Wahana Pendidikan*, *8*(11), 207–215

Alqarni, M., Gupta, B., & Zhang, X. (2020). A hybrid clustering approach based on k-means and differential evolution for efficient segmentation of big data. Future Generation Computer Systems, 105, 243-253.

Ananda, M. Risqi, Sandra, Nurul, Fadhila, Eka, Rahma, Alvia, & Nurbaiti, Nurbaiti. (2024). Data Mining dalam Perusahaan PT Indofood Lubuk Pakam. *Comit: Communication, Information and Technology Journal*, *2*(1), 108–119.

Huda, Miftahul, & Kom, M. (2019). *Algoritma Data Mining: Analisis Data Dengan Komputer*. bisakimia.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

Hasan, M. K., Al-Mamun, M. A., & Rahman, M. (2018). A comparative study of cluster validity indices for pattern recognition in data mining. Journal of King Saud University-Computer and Information Sciences, 30(2), 135-149.

Jadhav, S., & Parab, S. (2018). Clustering analysis using k-means algorithm and elbow

method. International Journal of Advanced Research in Computer Engineering & Technology, 7(1), 18-23.

Liao, C. W., Chen, Y. T., Chen, H. H., & Chen, H. (2021). Customer segmentation for commercial banks using K-means algorithm. Journal of Industrial and Production Engineering, 38(1), 25-37. doi: 10.1080/21681015.2020.1855630.

Siregar, Amril Mutoi, Kom, S., Puspabhuana, M. Kom D. A. N. Adam, Kom, S., & Kom, M. (2017). *Data Mining: Pengolahan Data Menjadi Informasi dengan RapidMiner*. CV Kekata Group.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Ye, Q., Lu, S., & Hu, Y. (2020). Customer segmentation and analysis of e-commerce big data based on K-means algorithm. Journal of Physics: Conference Series, 1616(5), 052035. doi: 10.1088/1742-6596/1616/5/052035.