Analisis Segmentasi Pelanggan berdasarkan Pendapatan dan Pengeluaran Menggunakan Algoritma Clustering K-Means

Customer Segmentation Analysis based on Income and Spending Using K-Means Clustering Algorithm

Chyntia Priseillia, Felix Samuel Leo, Nelson Saputra, Shyfa Ariesta Rustian, Tasya Chairunisa
Sistem Informasi, Fakultas Teknik Informatika, Universitas Multimedia Nusantara,
chyntia.priseillia@student.umn.ac.id, felix.samuel@student.umn.ac.id, nelson.saputra@student.umn.ac.id, shyfa.ariesta@student.umn.ac.id, tasya.chairunisa@student.umn.ac.id

***Abstrak*— Penelitian ini berfokus pada analisis segmentasi pelanggan berdasarkan pendapatan dan pengeluaran menggunakan algoritma clustering K-Means dan melalui proses metodologi CRISP-DM. Tujuan utama dari penelitian ini adalah untuk mengetahui hasil segmentasi pelanggan menggunakan metode *clustering K-Means* berdasarkan data pendapatan dan pemasukan pelanggan dari suatu supermarket yang berguna sebagai patokan data pelanggan sehingga supermarket bisa menggunakannya sebagai bahan acuan pengambilan keputusan pada bisnis ini. Penelitian ini menyimpulkan bahwa algoritma K-Means lebih efektif dalam mengidentifikasi segmen pelanggan yang berbeda berdasarkan pendapat dan pengeluaran, serta memberikan dasar yang kuat untuk pengambilan keputusan bisnis yang lebih tepat sasaran.**

Kata kunci: Segmentasi Pelanggan, Pendapatan, Pengeluaran, K-Means, Clustering, Strategi Pemasaran, CRISP-DM.

I. INTRODUCTION

Seiring dengan berkembangnya era globalisasi saat ini, persaingan di pasar bisnis semakin meningkat. Oleh karena itu setiap perusahaan wajib untuk memiliki keunggulan untuk bersaing dengan berbagai pesaing di dunia bisnis. Supermarket merupakan salah satu proyek bisnis perusahaan yang saat ini berkembang sangat pesat di industri sekarang ini. [1] Pada era industri sekarang, masyarakat sangat memerlukan berbelanja ke pasar atau sebuah supermarket untuk memenuhi kebutuhan sehari-hari. Pada era modern ini, manusia hanya mengolah bisa mengolah bahan mentah untuk dijadikan makanan sehari-hari. Supermarket menjadi salah satu bentuk bisnis retail yang menawarkan berbagai macam produk kepada pelanggan mereka. Dalam era persaingan saat ini supermarket memerlukan sebuah acuan untuk mempertahankan pelanggan yang ada agar selalu memilih supermarket tersebut. [2] Dengan menawarkan diskon, pelayanan prima, maupun inovasi produk, pelanggan menjadi lebih setia kepada supermarket tersebut karena sesuai dengan cakupan pendapatan yang ada. Supermarket bisa memanfaatkan data-data pelanggan seperti pendapatan dan pengeluaran pelanggan untuk mengadakan sebuah diskon agar pelanggan tetap kembali pada supermarketnya.

Dengan menggunakan *Data Mining* untuk mendapatkan informasi yang dapat dimanfaatkan

dan sangat bernilai sudah menjadi rahasia umum pada saat ini. [2] Pemanfaatan data mining pada proses bisnis retail dapat menggali informasi yang sebelumnya tidak diketahui dan dipahami akan menjadi informasi yang sangat berguna sebagai acuan perkembangan bisnis ini. Pemanfaatan data mining untuk mengetahui proses pelanggan melakukan pembelian produk akan sangat berguna untuk perkembangan bisnis retail supermarket menjadi lebih baik lagi.

Penelitian ini menggunakan metode *clustering* untuk mengetahui segmentasi pelanggan berdasarkan dari pendapatan dan pengeluaran pelanggan pada suatu supermarket.. *Clustering* merupakan proses pengelompokan titik data menjadi dua atau lebih kelompok berdasarkan suatu kemiripan dan kesamaan dengan kelompok lainnya [3]. Melakukan *clustering customer* berguna untuk mengidentifikasi tingkah laku pelanggan berdasarkan pendapatan dan pengeluaran yang dilakukan oleh pelanggan. Metode *clustering* memiliki keunggulan yang baik untuk mengetahui tingkah laku pelanggan maka dari itu penelitian ini menggunakan teknik *clustering* sebagai metode untuk mengetahui keakuratan yang tinggi berdasarkan segmentasi pelanggan.

Clustering dibagi menjadi 2 metode secara umum yaitu *Hierarchical clustering,* dan *Partitional clustering.* Pada penelitian ini dilakukan *partitional clustering* yaitu *k-medoids* dan *k-means* karena kedua metode ini merupakan metode yang paling sederhana dan umum untuk digunakan dan cocok untuk mengatasi data yang dalam jumlah besar [4]. Berdasarkan penelitian Fatimah., dkk (2021) mengenai "Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit menggunakan metode *K-Means Clustering*" didapatkan bahwa metode *K-Means* merupakan metode terbaik untuk melakukan *clustering* dibandingkan dengan metode lainnya seperti *DBSCAN, GMM,* ataupun *Agglomerative Clustering* [5].

Penelitian ini bertujuan untuk mengetahui tingkah laku serta kebutuhan pelanggan yang diharapkan dapat membantu perkembangan supermarket agar lebih mudah menjaga kualitas produk serta kesetiaan pelanggan.

Penelitian ini tidak hanya memberikan kontribusi praktis bagi pengambilan keputusan dalam strategi pemasaran, tetapi juga menambah literatur akademis mengenai penerapan algoritma K-Means dalam segmentasi pelanggan. Dengan demikian, penelitian ini diharapkan dapat menjadi referensi bagi praktisi bisnis dan peneliti di bidang pemasaran dan data analytics.

II. Studi Literatur

A. Clustering

Clustering merupakan metode mengumpulkan objek ke dalam masing-masing kelompok yang memiliki kesamaan menggunakan teknik *unsupervised learning,* salah satu teknik dalam *Machine learning algorithm.* Pada metode ini suatu objek yang memiliki karakteristik atau atribut yang sama akan dimasukkan kedalam satu kelompok cluster. Objek yang ada di dalam satu *cluster* akan memiliki karakteristik yang sama dan jika ada perbedaan maka akan sangat minimum [6].

B. K-Medoids

K-Medoids merupakan algoritma untuk menentukan *Medoids* dalam suatu *cluster* yang menjadi titik pusat sebuah *cluster.* Metode ini berbasis objek yang representative yang dapat memilih objek sebenarnya dalam mempresentasikan *cluster* daripada mengambil nilai rata-rata sebuah objek sebagai titik referensi [7].

C. K-Means

K-Means merupakan salah satu algoritma *unsupervised learning* yang paling sederhana dan sering digunakan untuk memecahkan masalah pengelompokkan yang ada. *K-Means* biasa digunakan untuk mengelompokkan data kedalam data *clustering* yang bertujuan agar meminimalkan fungsi dan variasi suatu objek yang ditentukan oleh *clustering*. Berikut adalah persamaan yang digunakan saat melakukan metode *K-Means*.

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

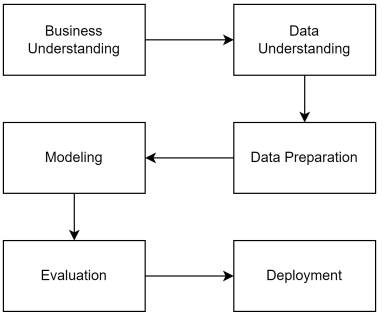
Dimana X^i = Himpunan titik dengan I =1, 2, 3, …, n dan dikelompokkan kedalam *cluster* dari satu set *cluster* yang diberikan sebagai ck dengan k = 1, 2, 3, …, k [7].

D. CRISP-DM

Metode *CRISP-DM* atau *Cross Industry Standard Process for Data Mining* merupakan sebuah metode penelitian data analisis yang bertujuan untuk memberikan tahapan dalam proses pengumpulan data. Dalam artikel oleh Navisa, et.al dikatakan bahwa proses tahapan yang ada di dalam *CRISP-DM* ada 6 tahapan yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment* [11]. Tahapan yang paling krusial dalam penerapan *CRISP-DM* adalah tahap *business understanding* yang memfokuskan pemahaman tujuan penelitian berdasarkan perspektif bisnis untuk melihat perbedaan metode *K-Medoids* dan *K-Means* terhadap pengeluaran dan pemasukan data *customer.* Tahapan berikutnya adalah tahapan *data understanding* yang merupakan tahapan untuk mengumpulkan data awal untuk kemudian diidentifikasikan dan dieksplorasi datanya.

III. METHODOLOGY

Berikut merupakan beberapa tahap yang dilakukan dalam penelitian ini:



Gambar 1. Alur Penelitian

Dalam gambar 1 dijelaskan tahapan-tahapan CRISP-DM yang dilakukan oleh peneliti saat melakukan penelitian ini untuk mencari segmentasi pelanggan berdasarkan data spending dan income menggunakan metode *K-Means.* Berikut merupakan penjabaran dari masing-masing tahapan.

3.1 Business Understanding

Pada tahap ini, penting untuk memahami tujuan bisnis di balik pengumpulan data yang ada. Analisis data pelanggan ini dimaksudkan untuk memberikan pemahaman yang berharga bagi perusahaan dalam meningkatkan layanan pelanggan, meningkatkan kepuasan pelanggan, dan mengoptimalkan strategi pemasaran. Selain itu, tujuan lainnya adalah untuk melakukan segmentasi pelanggan berdasarkan beberapa atribut seperti pendapatan, usia, dan frekuensi pembelian. Ini akan memungkinkan perusahaan untuk menyesuaikan penawaran dan promosi secara lebih efektif kepada setiap segmen pelanggan.

Dataset customer_data.csv menyediakan informasi tentang nama pelanggan, usia, jenis kelamin, pendidikan, pendapatan, dan frekuensi pembelian. Dengan menggunakan atribut-atribut ini, tujuan utama penelitian adalah untuk mengelompokkan

pelanggan ke dalam segmen yang homogen berdasarkan perilaku belanja mereka. Mengingat dataset ini terdiri dari 1000 baris dan 6 kolom, analisis lebih lanjut akan dilakukan untuk memahami pola dan tren di antara pelanggan. Hasil analisis ini akan menjadi dasar untuk pengambilan keputusan dalam pengembangan strategi pemasaran dan manajemen pelanggan.

3.2 Data Understanding

Pada tahap ini dilakukan pemahaman yang lebih mendalam terhadap dataset customer_data.csv. Data ini akan di import ke dalam lingkungan kerja dengan menggunakan library *pandas*. Proses ini memungkinkan peneliti untuk membaca dan memeriksa struktur dataset, termasuk jumlah baris dan kolom, serta tipe data tiap kolomnya. Setelah proses import dilakukan, langkah selanjutnya adalah memeriksa informasi data mengenai dataset menggunakan fungsi cust.info().

```
cust.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   name                1000 non-null   object
 1   age                 1000 non-null   int64
 2   gender              1000 non-null   object
 3   education           1000 non-null   object
 4   income              1000 non-null   int64
 5   country             1000 non-null   object
 6   purchase_frequency  1000 non-null   float64
 7   spending            1000 non-null   float64
dtypes: float64(2), int64(2), object(4)
memory usage: 62.6+ KB
```

Gambar 2. Struktur dataset

Informasi dari fungsi tersebut mencakup jumlah baris dan kolom, serta tipe data dari setiap kolom. Hal ini penting untuk memahami karakteristik data yang akan digunakan dalam analisis selanjutnya.

```
cust.describe()
```

	age	income	purchase_frequency	spending
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	41.754000	59277.852000	0.554600	9613.296835
std	13.778582	23258.377128	0.284675	5484.707210
min	18.000000	20031.000000	0.100000	611.985000
25%	30.000000	38825.500000	0.300000	5020.425000
50%	42.000000	58972.000000	0.600000	9430.395000
75%	54.000000	79114.000000	0.800000	13645.507500
max	65.000000	99780.000000	1.000000	25546.500000

Gambar 3. Deskripsi Statistik dataset

Pada gambar di atas memberikan hasil ringkasan statistik deskriptif dari kolom-kolom numerik dalam data frame. Kolom-kolom yang termasuk dalam kolom Age, Income, Purchase_frequency, dan Spending. Analisis deskriptif ini memberikan gambaran awal mengenai distribusi dan karakteristik data pelanggan yang akan membantu dalam proses segmentasi dan analisis lebih lanjut.

```
cust.dtypes
```

```
name                 object
age                   int64
gender               object
education            object
income                int64
country              object
purchase_frequency  float64
spending            float64
dtype: object
```

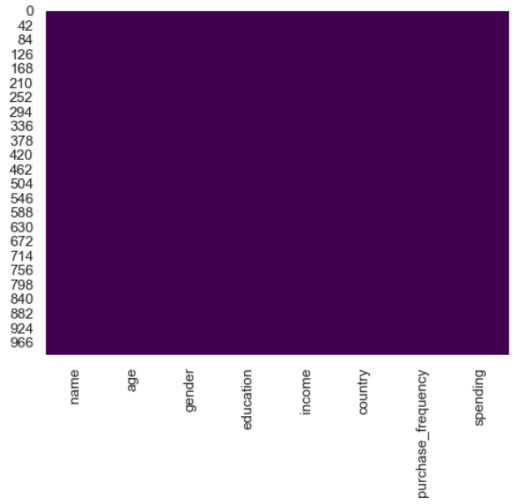
Gambar 4. Tipe Kolom dataset

Pada gambar di atas memberikan informasi mengenai tipe data dari setiap kolom dalam data frame. Kolom dengan tipe Object berisikan data kategorikal atau teks, sedangkan kolom dengan tipe int64 dan float64 berisikan data numerik yang bisa digunakan untuk melakukan analisis statistik, dll.

3.3 Data Preparation

Langkah selanjutnya yaitu proses *data preparation* dengan yang pertama adalah melakukan pemeriksaan keberadaan nilai yang hilang (*missing values*) dalam dataset menggunakan fungsi isnull() dan visualisasi heatmap untuk memastikan kelengkapan dan keandalan data.

```
sns.heatmap(cust.isnull(), cbar=False, cmap='viridis')
plt.show()
```



Gambar 3. Visualisasi heatmap

```
cust.isnull().any().any()
```

```
False
```

```
cust.isna().sum()
```

```
name                0
age                 0
gender              0
education           0
income              0
country             0
purchase_frequency  0
spending            0
dtype: int64
```

Gambar 4. Fungsi isnull()

```
cust.duplicated().sum()
```

```
0
```

Gambar 5. Cek duplikat

Setelah dilakukan pemeriksaan missing values, langkah selanjutnya yaitu mengecek kolom yang mempunyai duplikat dengan menggunakan fungsi duplicated() dan pada dataset disini tidak terdapat duplikat di tiap kolom yang ada.

```
cust.drop(['name'], axis=1, inplace=True)
cust.head()
```

	age	gender	education	income	country	purchase_frequency	spending
0	42	Female	High School	53936	Slovenia	0.9	13227.120
1	49	Female	Master	82468	Aruba	0.6	12674.040
2	55	Male	Bachelor	56941	Cyprus	0.3	5354.115
3	24	Female	Bachelor	60651	Palau	0.2	2606.510
4	64	Male	Master	81884	Zambia	0.9	18984.780

Gambar 6. Menghapus kolom

Fungsi dari drop() disini untuk menghapus kolom data yang tidak diperlukan. Pada dataset disini, kolom yang tidak diperlukan adalah “name”.

```
cust.gender.value_counts()
```

```
gender
Male      501
Female    499
Name: count, dtype: int64
```

```
cust.education.value_counts()
```

```
education
Bachelor      271
PhD           248
High School   245
Master        236
Name: count, dtype: int64
```

```
cust.country.value_counts()
```

```
country
Congo                12
Palau                11
Slovenia             10
Algeria               9
Ukraine               9
                     ..
Equatorial Guinea     1
Solomon Islands       1
Niger                 1
Botswana              1
Sudan                 1
Name: count, Length: 239, dtype: int64
```

Gambar 7. Value data

Pada tahap ini dilakukan pemeriksaan distribusi data *gender, education, dan country* untuk memahami karakteristik dasar dari dataset. Setelah itu, dataset dimodifikasi dengan mengubah data kategorikal menjadi numerik, menghapus kolom yang tidak diperlukan, dan memastikan format data sesuai untuk analisis selanjutnya. Langkah-langkah ini penting untuk meningkatkan kualitas dan keandalan hasil analisis clustering.

Pada tahap data preparation, dilakukan juga pemilihan variabel yang akan dianalisis lebih lanjut. Proses ini sangat penting untuk memastikan bahwa hanya variabel yang relevan dan memberikan kontribusi signifikan terhadap analisis clustering yang dipertahankan untuk tahapan selanjutnya.

```
num_col = cust.select_dtypes(include=np.number)
num_col
```

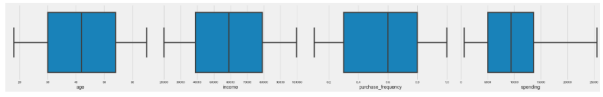
	age	income	purchase_frequency	spending
0	42	53936	0.9	13227.120
1	49	82468	0.6	12674.040
2	55	56941	0.3	5354.115
3	24	60651	0.2	2606.510
4	64	81884	0.9	18984.780
...	...	...	...	...
995	42	98170	0.7	17435.950
996	39	59174	0.9	11662.830
997	50	59808	0.4	7196.160
998	50	58272	0.7	13939.520
999	38	32512	0.5	8312.800

1000 rows × 4 columns

Gambar 8. Memilih DataTypes

Pada Tahap ini dilakukan pemilihan data types yang berkategori numeric, kolom yang termasuk ke dalam kategori numerik adalah Age, Income, Purchase_frequency, dan Spending. Data types ini dapat digunakan untuk analisis lebih lanjut untuk proses modeling.

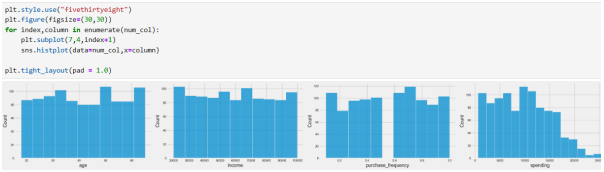
```
plt.style.use("fivethirtyeight")
plt.figure(figsize=(30,30))
for index,column in enumerate(num_col):
    plt.subplot(7,4,index+1)
    sns.boxplot(data=num_col,x=column)

plt.tight_layout(pad = 1.0)
```



Gambar 9. Cek Outlier

Pada tahap ini dilakukan pengecekkan terhadap outlier agar pada tahap modeling tidak terdapat perbedaan yang signifikan. Bisa dilihat bahwa tidak terdapat outlier dari 4 kolom yang dipilih pada proses pemilihan data types.

```
plt.style.use("fivethirtyeight")
plt.figure(figsize=(30,30))
for index,column in enumerate(num_col):
    plt.subplot(7,4,index+1)
    sns.histplot(data=num_col,x=column)

plt.tight_layout(pad = 1.0)
```

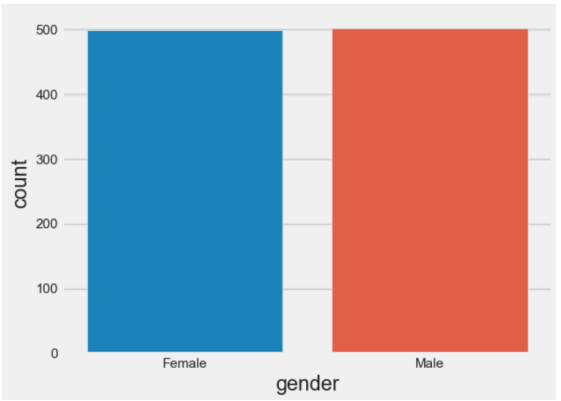


Gambar 10. Visualisasi histogram

Histogram ini memberikan gambaran visual yang jelas tentang karakteristik dasar dari variabel numerik dalam dataset. Dengan informasi ini, peneliti dapat membuat keputusan yang lebih baik terkait persiapan data dan teknik analisis yang akan digunakan selanjutnya.

Dalam tahap ini, peneliti melakukan pemahaman terhadap distribusi data dengan menggunakan visualisasi seperti histogram dan boxplot. Visualisasi ini membantu untuk mengidentifikasi pola dan penyebaran data, serta menemukan potensi adanya data outlier yang perlu diperhatikan dalam analisis lanjutan. Berikut merupakan hasil visualisasi data yang ada:

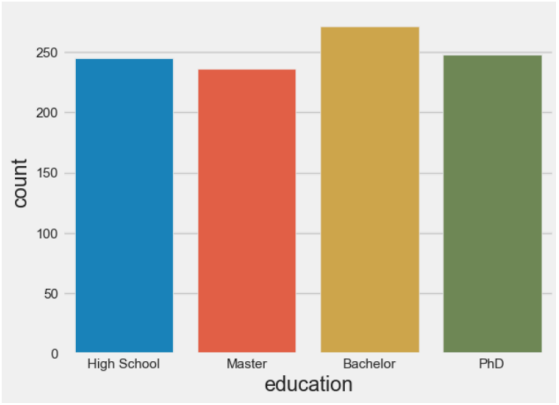
```
sns.countplot(x='gender', data=cust)
plt.show()
```



Gambar 11. Distribusi gender

Dengan visualisasi ini, kita dapat dengan cepat melihat bahwa dataset memiliki distribusi gender yang seimbang, yang penting dalam memastikan bahwa analisis selanjutnya tidak dipengaruhi oleh ketidakseimbangan data pada kategori gender.

```
sns.countplot(x='education', data=cust)
plt.show()
```

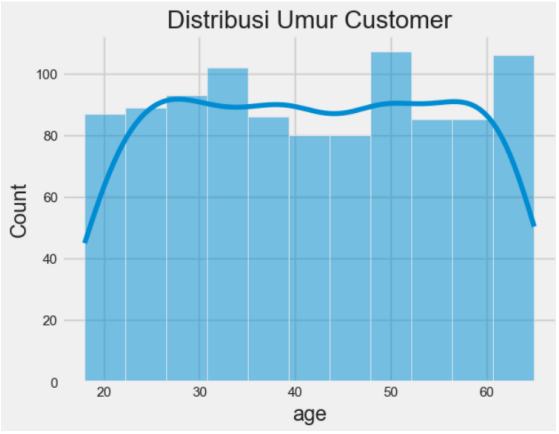


Gambar 12. Distribusi education

Dengan visualisasi ini, kita dapat dengan cepat melihat bahwa dataset memiliki distribusi education. Dilihat dari gambar di atas bahwa yang tertinggi yaitu Bachelor dan yang terendah yaitu Master.

```
sns.histplot(cust.age, kde=True)
plt.title("Distribusi Umur Customer")
```

```
Text(0.5, 1.0, 'Distribusi Umur Customer')
```

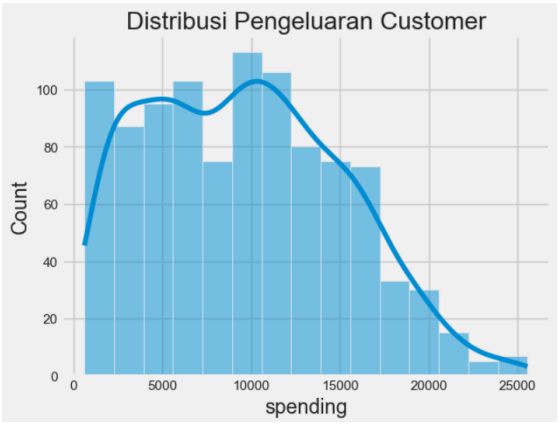


Gambar 13. Visualisasi umur customer

Terlihat pada gambar diatas terdapat visualisasi data distribusi umur customer. Digunakan visualisasi histogram karena dapat melihat dengan jelas penyebaran data yang terbagi kedalam beberapa nilai tertentu.

```
sns.histplot(cust.spending, kde=True)
plt.title("Distribusi Pengeluaran Customer")
```

```
Text(0.5, 1.0, 'Distribusi Pengeluaran Customer')
```

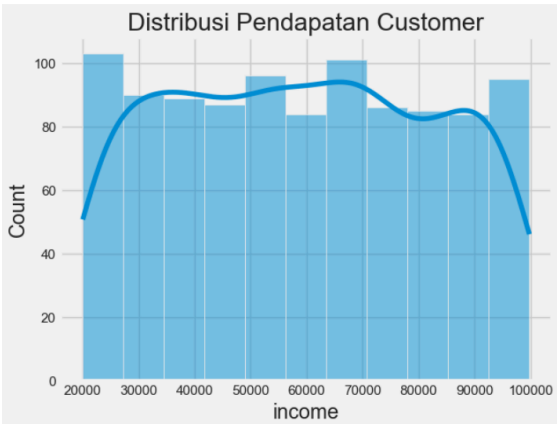


Gambar 14. Visualisasi pengeluaran customer

Terlihat pada gambar diatas terdapat visualisasi data distribusi pengeluaran customer. Digunakan visualisasi histogram karena dapat melihat dengan jelas penyebaran data yang terbagi kedalam beberapa nilai tertentu.

```
sns.histplot(cust.income, kde=True)
plt.title("Distribusi Pendapatan Customer")
```

```
Text(0.5, 1.0, 'Distribusi Pendapatan Customer')
```



Gambar 15. Visualisasi pendapatan customer

Terlihat pada gambar diatas terdapat visualisasi data distribusi pengeluaran dan pendapatan customer. Digunakan visualisasi histogram karena dapat melihat dengan jelas penyebaran data yang terbagi kedalam beberapa nilai tertentu.

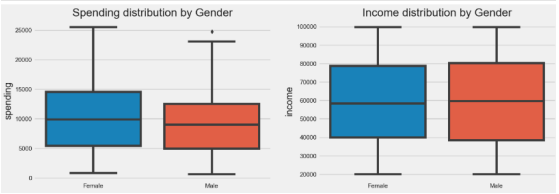
Selain itu dilakukan juga visualisasi scatter plot seperti gambar di bawah ini.

```
sns.scatterplot(data=cust, x='income', y='spending', hue='gender')

plt.title('Spending vs Income')
plt.xticks(rotation = 25)
plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
plt.show()
```

Gambar 16. Scatterplot spending vs income

Visualisasi diatas digunakan untuk membantu memvisualisasikan hubungan antara variabel spending dan income berdasarkan gemder dapat dilihat bahwa penyebaran data spending didominasi oleh gender wanita sedangkan pendapatan didominasi oleh gender pria.

```
plt.figure(figsize=(15,5))

plt.subplot(1,2,1)
sns.boxplot(x='gender', y='spending', data=cust)
plt.xlabel('')
plt.title('Spending distribution by Gender')

plt.subplot(1,2,2)
sns.boxplot(x='gender', y='income', data=cust)
plt.xlabel('')
plt.title('Income distribution by Gender')
plt.show()
```



Gambar 17. Visualisasi boxplot

Terlihat pada gambar diatas terdapat visualisasi data distribusi pengeluaran dan pendapatan customer. Diperjelas dalam visualisasi diatas bahwa pelanggan wanita memiliki pengeluaran lebih besar dibandingkan dengan pelanggan pria, sedangkan pendapatan pelanggan pria lebih tinggi dari pelanggan wanita.

```
cust_copy = cust
cust_copy = pd.get_dummies(cust_copy, columns=['gender'])
cust_copy.drop(['country','education'], axis=1, inplace=True)
cust_copy.head()
```

	age	income	purchase_frequency	spending	gender_Female	gender_Male
0	42	53936	0.9	13227.120	True	False
1	49	82468	0.6	12674.040	True	False
2	55	56941	0.3	5354.115	False	True
3	24	60651	0.2	2606.510	True	False
4	64	81884	0.9	18984.780	False	True

Gambar 18. Encoding data cust_copy

Pada gambar diatas dilakukan encoding data customer sehingga data yang masih bersifat objek seperti gender diubah menjadi encoding kolom gender baru. Selain itu juga dilakukan penghapusan kolom yang tidak dipakai yaitu ‘country’ dan ‘education’.

```
from sklearn.preprocessing import MinMaxScaler

cust_copy[cust_copy.columns] = MinMaxScaler().fit_transform(cust_copy)
```

```
cust_copy = cust_copy.reset_index(drop=True)
cust_copy.head()
```

	age	income	purchase_frequency	spending	gender_Female	gender_Male
0	0.510638	0.425146	0.888889	0.505931	1.0	0.0
1	0.659574	0.782919	0.555556	0.483749	1.0	0.0
2	0.787234	0.462827	0.222222	0.190183	0.0	1.0
3	0.127660	0.509348	0.111111	0.079991	1.0	0.0
4	0.978723	0.775596	0.888889	0.736842	0.0	1.0

Gambar 19. Normalisasi data cust_copy

Normalisasi data dengan MinMaxScaler mengubah skala fitur-fitur dalam dataset ke rentang [0, 1], memastikan kontribusi setara dalam analisis atau model machine learning. Ini mencegah dominasi fitur berskala besar dan penting untuk algoritma clustering seperti K-Means yang bergantung pada jarak antar data poin.

Gambar 20. Uji coba metode K-Medoids

Gambar diatas merupakan visualisasi dengan menggunakan metode Elbow yang

berfungsi untuk menentukan jumlah klaster pada K-Medoids.

```
cust_copy['K-Medoids'] = clusters
cust_copy.head()
```

	age	income	purchase_frequency	spending	gender_Female	gender_Male	K-Medoids
0	0.510638	0.425146	0.888889	0.505931	1.0	0.0	4
1	0.659574	0.782919	0.555556	0.483749	1.0	0.0	4
2	0.787234	0.462827	0.222222	0.190183	0.0	1.0	1
3	0.127660	0.509348	0.111111	0.079991	1.0	0.0	4
4	0.978723	0.775596	0.888889	0.736842	0.0	1.0	3

Gambar 21. Hasil Uji coba metode K-Medoids

Pada gambar diatas memberikan hasil clustering dari k-medoids dan dibagi menjadi 5 cluster sesuai dengan hasil elbow method. Setiap nomor klaster dalam kolom 'K-Medoids' menunjukkan kelompok yang berbeda hasil dari clustering oleh k-medoids.

Gambar 22. Uji coba metode K-Means

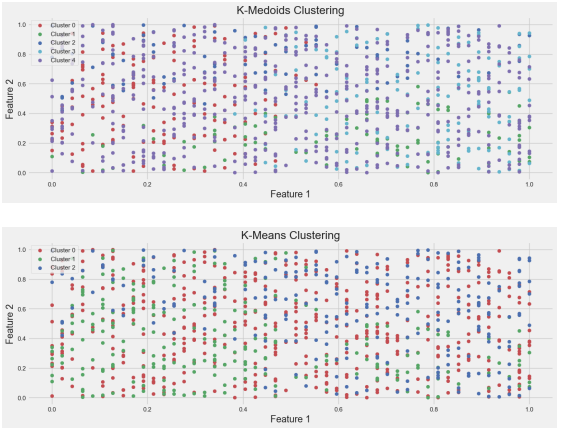
Gambar diatas merupakan visualisasi dengan menggunakan metode Elbow yang berfungsi untuk menentukan jumlah klaster pada K-Means.

```
cust_copy["K-Means"] = labels
cust_copy.head(5)
```

age	income	purchase_frequency	spending	gender_Female	gender_Male	K-Medoids	K-Means
0.510638	0.425146	0.888889	0.505931	1.0	0.0	4	0
0.659574	0.782919	0.555556	0.483749	1.0	0.0	4	0
0.787234	0.462827	0.222222	0.190183	0.0	1.0	1	1
0.127660	0.509348	0.111111	0.079991	1.0	0.0	4	0
0.978723	0.775596	0.888889	0.736842	0.0	1.0	3	2

Gambar 22. Hasil Uji coba metode K-Means

Pada gambar diatas memberikan hasil clustering dari k-means dan dibagi menjadi 3 cluster sesuai dengan hasil elbow method. Setiap nomor klaster dalam kolom 'K-Means' menunjukkan kelompok yang berbeda hasil dari clustering oleh k-means.



Gambar 23. Hasil plotting kedua metode

Gambar diatas merupakan visualisasi dari hasil clustering K-Medoids dan K-Means dengan menggunakan scatter plot.

Evaluasi K-Medoids dan K-Means menggunakan Silhoutte Score

```
from sklearn.metrics import silhouette_score

# Hitung Silhouette Score untuk K-medoids
silhouette_kmedoids = silhouette_score(cust_copy, clusters)
print(f"Silhouette Score untuk K-medoids: {silhouette_kmedoids}")

# Hitung Silhouette Score untuk K-means
silhouette_kmeans = silhouette_score(cust_copy, labels)
print(f"Silhouette Score untuk K-means: {silhouette_kmeans}")
```

```
Silhouette Score untuk K-medoids: 0.6753118812505097
Silhouette Score untuk K-means: 0.6880564426689733
```

Gambar 24. Hasil evaluasi kedua metode

Dari hasil evaluasi perbandingan metode *K-Medoids* dan *K-Means* didapatkan bahwa metode *K-Means* lebih cocok digunakan untuk model segmentasi pelanggan pada data ini dikarenakan melihat skor dari *silhouette* menunjukkan bahwa nilai untuk *K-Means* lebih besar dibandingkan dengan *K-Medoids*.

3.4 Modelling

Pada tahapan ini dilakukan *modelling* data dengan menggunakan algoritma yang sudah dipilih yaitu *K-Means*. Metode *K-Means* dipilih karena setelah dilakukan evaluasi model sementara pada tahapan data preparation didapatkan bahwa metode *K-Means* menjadi metode yang lebih sesuai untuk melakukan *clustering* segmentasi pelanggan untuk dataset ini.

Pada pemodelan data, data yang sudah siap disimpan kedalam data frame ‘cust_copy’ yang akan digunakan untuk pemodelan data pada metode *K-Means*. Dataframe ‘cust_copy’ sudah dilakukan standarisasi untuk permulaan pemodelan menggunakan *MinMaxScaller*. Fungsi standarisasi *MinMaxScaller* adalah untuk mengubah data yang ada sehingga setiap fiturnya memiliki rentang tertentu seperti 0 dan 1 dengan dilakukan rentang data (*scaling)* dan memindahkan data (*shifting*). Berikut adalah dataframe yang digunakan untuk pemodelan data segmentasi pelanggan.

```
cust_copy = cust_copy.reset_index(drop=True)
cust_copy.head()
```

	age	income	purchase_frequency	spending	gender_Female	gender_Male
0	0.510638	0.425146	0.888889	0.505931	1.0	0.0
1	0.659574	0.782919	0.555556	0.483749	1.0	0.0
2	0.787234	0.462827	0.222222	0.190183	0.0	1.0
3	0.127660	0.509348	0.111111	0.079991	1.0	0.0
4	0.978723	0.775596	0.888889	0.736842	0.0	1.0

Gambar 25. Dataframe cuts_copy

Setelah memiliki dataframe yang akan digunakan untuk melakukan modeling data menggunakan *K-Means,* pertama-tama akan dibuat objek PCA yang merupakan *Principal Component Analysis* yang berguna untuk mengurangi dimensi data, dengan mempertahankan sebanyak mungkin variabilitas dalam data yang ada. Objek ini akan disimpan kedalam fungsi ‘pca’ yang kemudian diterapkan kepada data dengan menyesuaikan ‘fit’ data frame ‘cust_copy’ dan disimpan kedalam fungsi ‘principalComponents’. Setelah itu akan dibuat dataframe baru berupa ‘principalDf’ yang berisi hasil fitting fungi ‘pca’ kedalam dataframe ‘cust_copy’. Berikut adalah codenya.

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)

principalComponents = pca.fit_transform(cust_copy)

principalDf = pd.DataFrame(data = principalComponents
             , columns = ['principal component 1', 'principal component
```

Gambar 26. Deklarasi fungsi PCA

Setelah mendeklarasikan PCA didapatkan hasil sebagai berikut:

Gambar 27. Visualisasi hasil PCA

Gambar diatas merupakan hasil visualisasi PCA dua dimensi yang menunjukkan proyeksi data pada dua komponen utama. Plot tersebut didapatkan bahwa data tersebar dalam dua dimensi utama yang ditangkap oleh PCA dan menjadi cluster dalam data ini.

Clustering K-Means

```
model = KMeans(n_clusters=4,random_state=42, n_init='auto')
model.fit(principalComponents)
```

```
KMeans(n_clusters=4, n_init='auto', random_state=42)
```

Gambar 28. Model Hasil PCA K-Means

Setelah melakukan deklarasi PCA kemudian disimpan kepada model untuk diterapkan dalam analisis PCA menggunakan *KMeans* dan disimpan kedalam dataframe ‘model’.

```
labels = model.labels_
```

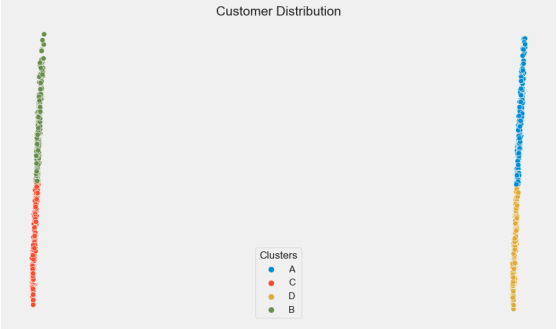
```
mapping = {0: 'A', 1: 'B', 2: 'C', 3: 'D'}

labels_mapping = [mapping[val] for val in labels]
```

```
cust['Clusters'] = labels_mapping
cust_copy['Clusters'] = labels_mapping
principalDf['Clusters'] = labels_mapping
```

Gambar 29. Mapping Clusters

Gambar diatas adalah pemberian label pada cluster yang ada yaitu label ‘A’ untuk cluster 0, label ‘B’ untuk cluster 1, label ‘C’ untuk cluster 3, dan label ‘D’ untuk cluster 4. Label ini kemudian diterapkan pada dataframe ‘cust’, ‘cust_copy’, dan ‘principalDf’. Setelah menyimpan label cluster data, kemuudian dilakukan visualisasi penyebaran cluster hasil segmentasi kedalam scatter plot yang masing-masing warna menggambarkan

masing-masing cluster yang ada sesuai dengan 4 label cluster ‘A’, ‘B’, ‘C’, dan ‘D’. Hasil dari scatter plot ini menunjukkan distribusii cluster pelanggan dalam dua dimensi utama yang dihasilkan oleh PCA. Visualisasi ini membantu mengidentifikasikan pola cluster dalam data pelanggan yang berguna untuk analisis lebih lanjut pada pengambilan keputusan dalam perusahaan. Berikut adalah hasil visualisasi model cluster:



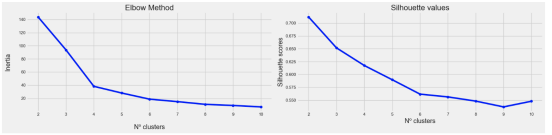
Gambar 30. Visualisasi Hasil Clusters

3.5 Evaluation

Setelah melakukan pemodelan data, model data tersebut akan di evaluasi menggunakan *elbow method* dan juga *silhouette score* yang merupakan sebuah metode pengujian model *cluster* yang dilakukan untuk mengetahui seberapa dekat relasi antara objek didalam sebuah *cluster. Elbow method* dilakukan untuk mengukur total jarak kuadrad tiap titik ke centroid terdekat, dengan inertia menunjukkan titik lebih dekat ke centroidnya sehingga cluster yang dibentuk akan menjadi lebih baik. Nilai yang terdapat pada *silhouette coefficient* ini berkisar dalam -1 hingga 1. Semakin nilai hasilnya mendekati angka 1 maka semakin baik model *clustering* data yang di uji. Sebaliknya semakin mendekati angka -1, maka semakin buruk model *clustering* data yang dilakukan [10]. Berikut merupakan hasil dari evaluasi model *K-Means:*

```
inertias = []
silhouette_scores = []

for k in range(2, 11):
    model = KMeans(n_clusters=k, random_state=42, n_init='auto')
    model.fit(principalComponents)
    inertias.append(model.inertia_)
    silhouette_scores.append(silhouette_score(principalComponents, model.labels_))
```

```
plt.style.use("fivethirtyeight")
fig,ax=plt.subplots(1,2,figsize=(20,5))
ax[0].plot(range(2,11), inertias,color='blue',marker='o',label='Elbow Method')
ax[0].set_xticks(range(2,11))
ax[0].set_xlabel("Nº clusters", labelpad=20)
ax[0].set_ylabel("Inertia", labelpad=20)
ax[0].set_title("Elbow Method")

ax[1].plot(range(2,11), silhouette_scores,color='blue',marker='o',label='Silhouette scores')
ax[1].set_xticks(range(2,11))
ax[1].set_title("Silhouette values")
ax[1].set_xlabel("Nº clusters")
ax[1].set_ylabel("Silhouette scores")
plt.tight_layout();
```



Gambar 31. Evaluasi model

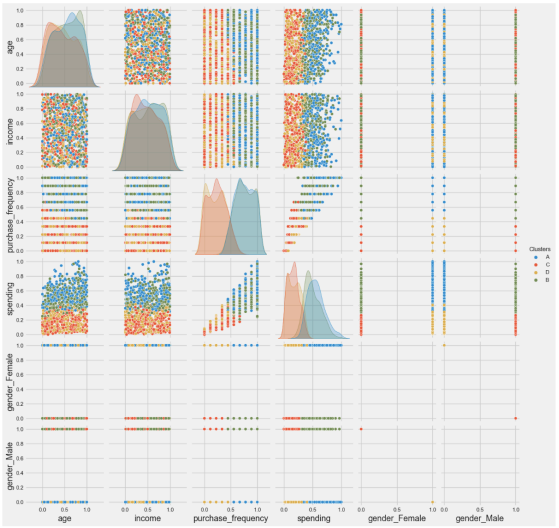
Dalam hasil evaluasi diatas, didapatkan grafik model yang telah dievaluasi menggunakan *elbow method* dan juga *silhouette score.* Dari grafik tersebut didapatkan wawasan berupa jumlah cluster optimal untuk model *K-Means* yang ada yaitu sebanyak 4 cluster, dengan evaluasi ditentukan bahwa model yang dibangun sudah merupakan model terbaik untuk melakukan segmentasi pelanggan dengan membagi cluster menjadi 4 bagian sesuai dengan evaluasi dalam *elbow method* dan juga *silhouette score.*

3.6 Deployment

Fase terakhir dalam tahapan *CRISP-DM* adalah *deployment,* dalam tahapan ini penelitian ini membatasi pelaksanaan hingga tahapan evaluasi saja dan tidak melanjutkan kedalam tahapan penyebaran atau *deployment*. Pembuatan artikel laporan ini ditulis setelah melakukan evaluasi dari penelitian perbandingan algoritma *K-Medoids* dan *K-Means* yang mana sudah termasuk kedalam penerapan hasil penelitian. Diharapkan penelitian dimasa yang akan datang akan menggunakan hasil ini untuk membuat penelitian yang lebih baik.

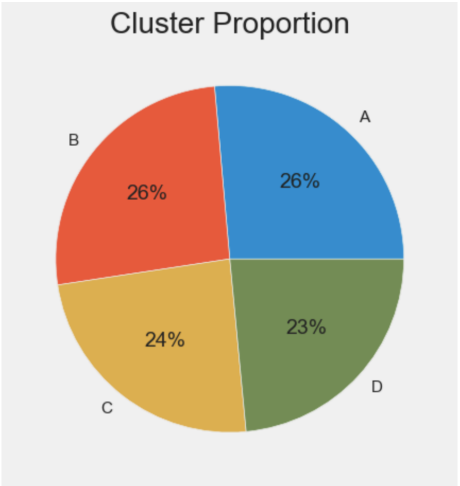
IV. Hasil dan Analisis

Dalam penelitian ini dilakukan juga pembahasan model yang digunakan untuk mencari segmentasi pelanggan berdasarkan hasil modeling dari metode *K-Means* yang dibuat, berikut adalah pembahasan hasil model segmentasi pelanggan.



Gambar 32. Visualisasi pairplot model

Pada gambar 32 divisualisasikan penyebaran data setelah dilakukan pemodelan, didapatkan bahwa beberapa data memiliki penyebaran data antar variabel yang teratur dan beberapa memiliki hubungan seperti variabel ‘age’ dengan ‘spending’ dan juga ‘income’.



Gambar 33. Visualisasi pie chart model

Visualisasi diatas merupakan perbandingan pembagian masing-masing cluster yang ada, dari hasil pie chart diatas didapatkan bahwa cluster terkecil adalah cluster ‘D’ yaitu 23% dan cluster ‘A’ dan ‘B’ memiliki besaran cluster yang sama yaitu 26%.

```
cust.groupby('Clusters')['spending'].mean()
```

```
Clusters
A    14685.959811
B    13045.364308
C     4730.815871
D     5124.604681
Name: spending, dtype: float64
```

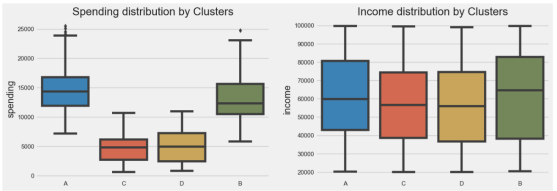
Gambar 34. Means Spending

```
cust.groupby('Clusters')['income'].mean()
```

```
Clusters
A    60976.356061
B    61650.961538
C    57219.639004
D    56854.940426
Name: income, dtype: float64
```

Gambar 35. Means Income

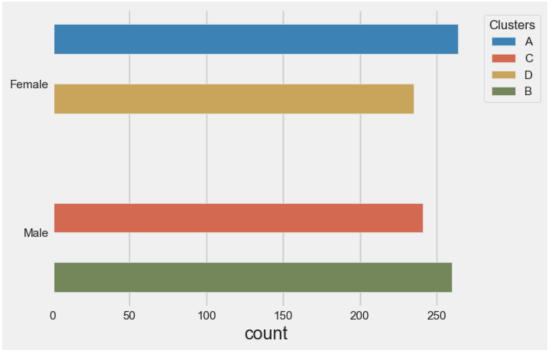
Pada gambar 34 dan 35 didapatkan rata-rata spending dan income dari masing-masing cluster yang ada. Pada rata-rata diatas terlihat bahwa cluster ‘A’ dan ‘B’ hanya memiliki sedikit perbedaan dalam rata-rata spending dan incomenya sedangkan dalam cluster ‘D’ terlihat bahwa cluster ini memiliki spending dan income yang paling kecil dari keempat cluster yang ada.



Gambar 36. Distribusi Spending dan Income per Cluster

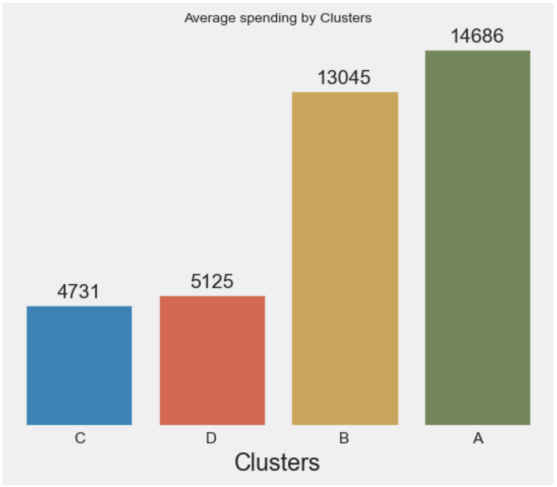
Distribusi spending dan income dari cluster diatas ditunjukkan dalam visualisasi boxplot dan didapatkan bahwa cluster ‘C’ memiliki spending yang paling kecil padalah income dari cluster ‘C’

termasuk besar dan juga income paling besar berada pada cluster ‘B’ yang juga memiliki distribusi spending paling banyak diantara cluster yang lainnya.



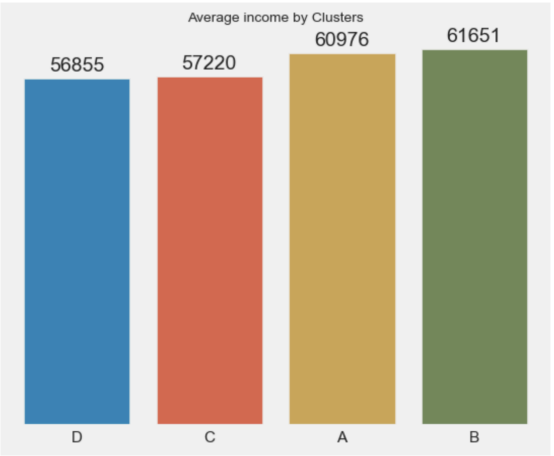
Gambar 37. Cluster by Gender

Dari hasil distribusi spending dan income dari cluster sebelumnya didapatkan bahwa cluster dengan spending tertinggi ada pada cluster ‘A’ dan ‘B’ dan penyebaran cluster tersebut rata dibagi dalam kedua gender. Porsi gender wanita dalam cluster ‘A’ lebih banyak melakukan spending dengan jumlah income yang sedikit dari pada cluster lain. Sedangkan porsi gender pria dalam cluster ‘B’ lebih banyak melakukan spending dengan jumlah income yang tinggi dari pada cluster lain.



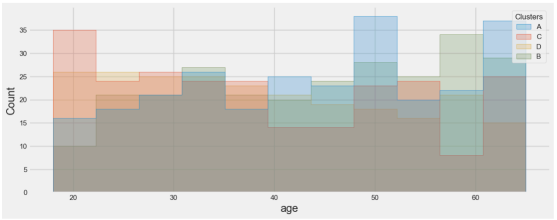
Gambar 38. Average Spending per Clusters

Bisa dilihat dalam gambar 38, didapatkan bahwa cluster ‘A’ memiliki jumlah spending yang paling tinggi dibandingkan dengan cluster lain. Sebaliknya, cluster ‘C’ memiliki tingkat spending paling rendah disini.



Gambar 39. Average Income per Clusters

Bisa dilihat dalam gambar 39, didapatkan bahwa cluster ‘B’ memiliki jumlah income yang paling tinggi dibandingkan dengan cluster lain. Sebaliknya, cluster ‘D’ memiliki tingkat income paling rendah disini.



Gambar 40. Distribusi umur per Clusters

Dalam gambar 40 ditampilkan visualisasi distribusi umur percluster. Didapatkan bahwa cluster ‘A’ didominasi oleh customer dengan umur yang lebih senior dari pada cluster ‘C’ dan dapat disimpulkan bahwa cluster customer ‘A’ lebih memiliki pekerjaan yang layak sehingga tingkat spendingnya paling tinffi dibandingkan dengan cluster lain.

Dari hasil analisis model didapatkan bahwa cluster ‘A’ didominasi oleh wanita dewasa yang spending score nya paling tinggi sehingga jika

perusahaan ingin meningkatkan penjualan maka ditargetkan kepada cluster ini. Selain cluster ‘A’, cluster ‘B’ juga menjadi cluster yang dapat diraih untuk melakukan spending lebih banyak karena cluster ini menempati tingkat kedua spending dengan tingkat pertama income, sehigga kemungkinan customer dalam cluster ini untuk menghabiskan uangnya lebih besar. Sebaliknya cluster ‘C’ dan ‘D’ didominasi oleh customer yang masih berusia muda sehingga incomenya masih belum terlalu besar untuk banyak melakukan spending.

V. Kesimpulan

Clustering data merupakan proses pengelompokan titik data menjadi dua atau lebih kelompok berdasarkan suatu kemiripan dan kesamaan dengan kelompok lainnya. Penelitian ini bertujuan untuk mengetahui hasil segmentasi pelanggan menggunakan metode *clustering K-Means* berdasarkan data pendapatan dan pemasukan pelanggan dari suatu supermarket yang berguna sebagai patokan data pelanggan sehingga supermarket bisa menggunakannya sebagai bahan acuan pengambilan keputusan pada bisnis ini.

Berdasarkan hasil dan pembahasan, didapatkan bahwa data ini lebih cocok digunakan *clustering* menggunakan metode *K-Means*. Dengan berpacu pada hasil evaluasi pada penelitian ini didapatkan bahwa metode *clustering K-Means* mendapatkan skor lebih baik dibandingkan dengan metode *clustering K-Medoids* dikarenakan data hasil evaluasi menjelaskan bahwa metode *K-Means* mendapatkan hasil sebesar 0.6880564426689733 yang mana lebih mendekati nilai 1. Dengan hasil skor yang mendekati nilai 1, metode *K-Means* menjadi metode *clustering* yang lebih baik daripada metode *K-Medoids*.

Dengan hasil yang didapatkan dilakukan modeling data menggunakan metode *K-Means* yang didapatkan bahwa pembagian cluster dalam data pelanggan ini dibagi menjadi 4 cluster yaitu ‘A’, ‘B’, ‘C’, dan ‘D’. Cluster ‘A’ didominasi oleh wanita dewasa yang spending score nya paling tinggi sehingga jika perusahaan ingin meningkatkan penjualan maka ditargetkan kepada cluster ini. Selain cluster ‘A’, cluster ‘B’ juga menjadi cluster yang dapat diraih untuk melakukan spending lebih banyak karena cluster ini menempati tingkat kedua spending dengan tingkat pertama income, sehigga kemungkinan customer dalam cluster ini untuk menghabiskan uangnya lebih besar.

Dari hasil pemodelan data ini diharapkan dapat membantu perusahaan supermarket dalam mengolah data pelanggan yang akan digunakan sebagai acuan pengambilan keputusan untuk keberlangsungan perusahaan lebih baik lagi kedepannya.

Dari hasil dan kesimpulan yang didapatkan peneliti memberikan beberapa saran untuk penelitian di masa depan yaitu dengan mencoba metode *clustering* yang lebih baik dibandingkan pada penelitian ini, dengan mengetahui lebih banyak metode maka diharapkan untuk lebih baik dalam pengembangan perusahaan.

References

[1] M. Uliyah and L. Sulistyawati, "Faktor-Faktor bauran pemasaran Yang Mempengaruhi loyalitas Pelanggan Berdasarkan Aspek-Aspeknya," *Reslaj : Religion Education Social Laa Roiba Journal,* vol. 4, no. 5, pp. 1238-1259, 2022.

[2] B. Y. Putra, F. Y. Azzahra and I. A. Erlanda, "Klasterisasi Pengunjung mall menggunakan algoritma K-means Berdasarkan pendapatan dan pengeluaran," *Jurnal Informatika dan Teknik Elektro Terapan,* vol. 11, no. 3s1, 2023.

[3] S. I. Attaqwa, A. A. Hanafi, H. Hakim, A. A. Sofyan and A. P. Sari, "Customer Clustering Menggunakan K-Means Agglomerative pada Pendapatan dan Pembelian daging," *Prosiding*

Seminar Nasional Informatika Bela Negara, vol. 3, no. 3, pp. 96-100, 2023.

[4] E. Irwansyah, "CLUSTERING," Binus University, [Online]. Available: https://socs.binus.ac.id/2017/03/09/clustering/. [Accessed 22 12 2023].

[5] F. D. Alhamdani, A. A. Dianti and Y. Azhar, "Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit menggunakan metode K-means clustering," *JISKA (Jurnal Informatika Sunan Kalijaga),* vol. 6, no. 2, pp. 70-77, 2021.

[6] M. Orisa, "Optimasi Cluster Pada algoritma k-means," *Prosiding SENIATI,* vol. 6, no. 2, pp. 430-437, 2022.

[7] M. Herviany, S. Putri Dilema, T. Nurhidayah and Kasini, "Perbandingan algoritma K-means Dan K-medoids untuk Pengelompokkan Daerah Rawan Tanah Longsor Pada provinsi Jawa Barat," *MALCOM: Indonesian Journal of Machine Learning and Computer Science,* vol. 1, no. 1, pp. 34-40, 2021.

[8] H. Fitriyah, E. M. Safitri, N. Muna, M. Khasanah, D. A. Aprilia and D. Nurdiansyah, "IMPLEMENTASI ALGORITMA CLUSTERING DENGAN MODIFIKASI METODE ELBOW UNTUK MENDUKUNG STRATEGI PEMERATAAN BANTUAN SOSIAL DI KABUPATEN BOJONEGORO," *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika,* vol. 4, no. 3, pp. 1598-1607, 2023.

[9] I. G. I. Sudipa, I. B. G. Sarasvananda, Hartatik, H. Prayitno, I. N. T. A. Putra, R. Darmawan, D. A. WP and Efitra, Teknik Visualisasi Data, Indonesia: PT. Sonpedia Publishing Indonesia, 2023.

[10] N. Nurhayati, "Pengujian Silhouette Coefficient," Blogger, 30 October 2018. [Online]. Available: https://nopi-en.blogspot.com/2018/11/pengujian-silhouette-coefficient.html. [Accessed 22 December 2023].

[11] S. Navisa, N. L. Hakim, and N. A. Nabilah, “Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM,” Jurnal Sistem Cerdas, vol. 4, no. 2, pp. 114–125, Aug. 2021, doi: 10.37396/jsc.v4i2.162.

Role Member

- **Chyntia Priseillia (00000070303)**
  1. Membuat code
  2. Membuat laporan bagian Modelling, Evaluation, Deployment, Hasil dan Analisis, serta Kesimpulan
- **Felix Samuel Leo (00000070094)**
  1. Membantu menyelesaikan laporan bagian methodology
- **Nelson Saputra (00000069095)**
  1. Membantu menjelaskan laporan bagian studi literatur dan melihat penelitian terdahulu agar dapat dimasukkan sebagai referensi
- **Shyfa Ariesta Rustian (00000071428)**
  1. Membantu menyelesaikan bagian data preparation di laporan
- **Tasya Chairunisa (00000071782)**
  1. Membuat Abstrak
  2. Membuat Pendahuluan
  3. Membantu menyelesaikan bagian data preparation (laporan)