Segmentasi Pelanggan Menggunakan Analisis Clustering KMeans

Indri Nur Sukmawati
Program Studi Sains Data, Fakultas Sains dan Teknologi
Universitas Koperasi Indonesia
Email: sukmawatiindri6@gmail.com

ABSTRAK

Penelitian ini bertujuan untuk mengetahui segmentasi pelanggan dari suatu perusahaan grosir. Menggunakan metode Kmeans Clustering, dataset yang digunakan dataset Customer Personalty Analysis denagn jumlah data 2240 data. Denagn dataset mencakup informasi demografi pelanggan seperti tahun lahir, status pernikahan, tingkat pendidikan dan jumlah anak serta income. Selain informasi pelanggan juga terdapat informasi produk yang meliputi wine, fruits, meat product, fish product, sweet product dan goldprods. Saluran pembelian mancakup deal purchases, web purchases, catalog purchases dan store purchases serta visist month. Dan juga ada respon kampanye. Proses analisis dimulai dari pengumpulan data, cleaning data, preprocessing data, dan clustering. Hasil analisis mengahsilkan emapat kluster yang diprofil berdasarkan struktur keluarga dan pendapatan/pengeluaran pelanggan.

Kata kunci: Kmeans Clustering, Segmentasi Pelanggan, Cleaning Data, Preprocessing Data.

1. PENDAHULUAN

Pemanfaatan teknologi sebagai alat bantu dalam mendukung aktivitas bisnis saat ini sangat memudahkan manusia dalam memperoleh informasi dengan cepat, tepat, dan akurat. Hal ini memungkinkan tujuan dari suatu pekerjaan dapat dicapai dengan lebih efektif dan efisien. Pemanfaatan teknologi juga dapat digunakan sebagai sarana promosi barang yang dimiliki suatu perusahaan. Semakin berkembangnya teknologi menimbulkan daya saing yang semakin kuat antar perusahaan, hal ini mendorong perusahaan untuk terus mencari cara inovatif dalam memahami perilaku konsumen guna merancang strategi pemasara yang lebih efektif. Salah satu pendekatan yang umum digunakan yaitu

segmentasi pelanggan. Dengan segmentasi yang tepat, perusahaan dapat menyesuaikan startegi pemasaran dan layanan mereka untuk memenuhi kebutuhan spesifik dari setiap segmen.

Salah satu metode yang sering digunakan dalam segmentasi pelanggan adalah analisis clustering. Clustering bertujuan untuk mengelompokkan data ke dalam kelompok (cluster) serupa sehingga dapat mengoptimalkan nilai dari suatu segmen. Teknik ini juga membantu perusahaan dalam memodelkan hubungan antara kecenderungan produk yang dibeli konsumen yang mungkin tidak diperhatikan oleh pihak perusahaan sebelumnya. Dengan mengidentifikasi segmen-segmen pelanggan yang berbeda, perusahaan dapat mengembangkan kampanye pemsaran yang lebih personal dan relevan, meningkatkan konversi dan kepuasan pelanggan.

Penelitian ini menerapkan Kmeans Clustering pada dataset pelanggan sebuah perusahaan grosir untuk membentuk segmentasi pelanggan. Dataset tersebut mencakup berbgai informasi seperi informasi pelanggan meliputi tahun lahir, pendidikan, status pernikahan, *income*, jumlah anak dan *recency.* Untuk data produk meliputi wine, fruits, meat products, fish product, sweet product, dan goldprods. Data tempat transaksi meliputi Web Purchases, Catalog Purchases, dan Visit Month. Dan juga data promosi mencakup Deal Purhases, AceptedCamp 1, AceptedCamp 2, AceptedCamp 3, AceptedCamp 4, AceptedCamp 4, Response. Dengan menggunakan Kmeans Clustering diharapakan dapat mengidentifikasi segmen-segmen pelanggan yang berbeda berdasarkan pola perilaku mereka.

2. TINJAUAN PUSTAKA

2.1 Segmentasi Pelanggan

Segmentasi adalah proses membagi pelanggan menjadi beberapa klaster dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran. Segmentasi pelanggan adalah salah satu langkah awal dalam membuat model bisnis.

Segmentasi pelanggan adalah proses memeriksa atribut pelanggan dan membuat kelompok berdasarkan bagaimana mereka berperilaku, siapa mereka, dan karakteristik spesifik mereka. Segmentasi pelanggan memungkinkan bisnis untuk menggunakan

pesan yang ditargetkan, dibandingkan menggunakan pendekatan yang bersifat universal, untuk mendorong hasil bisnis.

2.2 Data Mining

Data mining adalag metode yang memungkinkan para penggunanya untuk mengakses data yang besar dalam waktu yang relatif cepat. Atau dengan kata lain data mining merupakan suatu suatu alat dan aplikasi menggunakan analisis statistik pada data melalui suatu proses ekstraksi atau panggilan data dan informasi yang belum diketahui sebelumnya. Menurut (Pramudiono, 2007), Data mininng juga sering disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diektahui secara manual dari suatu kumpulan data.

2.3 Clustering

Analisis clustering merupakan sebuah analisis yang digunakan untuk mengelompokkan data yang tidak mempnya target/ label (*unsupervid*). Proses clustering merupakan proses pengelompokkan data berdasarkan kesamaan nilai fitur/ atribut. Selain mendapatkan hasil pengelompokkan data, dari penerapan analisis clustering akan didapatkan titik tengah dari data (*centroid*) dan cluster dengan jumlah anggota terbanyak.

2.4 Kmeans

Kmeans clustering suatu algoritma yang digunakan untuk mengelompokkan beberapa objek-objek berdasarkan atribut ke dalam beberapa k-kluster yang dimana jumlah k lebih kecil dari banyak objek biasa disebut *centroid*. Untuk menghitung jarak setiap kluster ialah dengan menemukan jarak yang paling dekat dari setiap data dengan *centroid* ini akan dilakukan *looping* hingga nilai cetroid stabil. Dengan rumus Kmeans Clustering sebagai berikut:

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad \ldots\ldots\ldots\ldots (1)$$

Dimana

$D_{(i,j)}$ = Jarak data ke I ke pusat cluster j

X_{ki} = Data ke I pada atribut data ke k

X_{kj} = Titik pusat ke j pada atribut ke k

2.5 Metode Elbow

Metode ini memberikan gambaran dengan memilih niali cluster kemudian menambahkan nilia cluster untuk digunakan sebagai model data dalam penentuan cluster terbaik. Hasil perhitungan presentase digunakan sebagai perbandingan antar cluster yang ditambahkan. Untuk mendapatkan perbandingan adalah menghitung SSE (Sum of Square Error) dari masing-masing nilai cluster. Karena semakin besar jumlah cluster K maka nilai SSE akan semakin kecil. Berikut adalah persamaan SSE dalam algoritma Kmeans:

$$\text{SSE} = \sum_{j=1}^{k} \sum x \in Cj \ \text{dist}(x, mj)^2 \ \ldots\ldots\ldots(2)$$

Dimana:

SSE = Jumlah error per-cluster

k = Jumlah nomer dari cluster

mj = Tititk data (Data Point)

$x \in Cj$ = Anggota titik data di cluster

3. METODE PENELITIAN

Merode penelitian yang digunakan pada penelitian ini adalah metode analisis clusering Kmenas dengan menggunakan bahasa pemrograman Pyhton dengan platform Google Colab. Data diperoleh dari website Kagle dengan jumlah data sebanyak 2240 data.

Dari dataset Customer Personalty Analysis akan diolah dengan Kmeans Clustering. Dataset mencakup informasi demografi pelanggan seperti tahun lahir, status pernikahan, tingkat pendidikan dan jumlah anak serta income. Selain informasi pelanggan juga terdapat informasi produk yang meliputi wine, fruits, meat product, fish product, sweet product dan goldprods. Saluran pembelian mancakup deal purchases, web purchases, catalog purchases dan store purchases serta visist month. Dan juga ada respon kampanye.

Langkah pertama yang dilakukan yaitu pengumpulan data, menggunakan dataset Customer Personalty Analysis yang mencakup atribut di atas. Langkah ke dua yaitu pra - pemrosesan, data yang akan di analisis dibersihkan dan di normalisasi untuk menghilangkan outlier dan memastikan konsisten. Langkah ke tiga yaitu implementasi

Kmrans Clustering, algoritma Kmeans diterapkan untuk mengidentifikasi kelompok-kelompok pelanggan.

4. HASIL PENELITIAN DAN PEMBAHASAN

Data yang digunakan untuk pemodelan clustering adalah Customer Personalty Analysis dengan jumlah data sebanyak 2240 data yang dimana isi nya mencakup informasi pelanggan, informasi produk dari grosir tersebut, jenis kampanye dan lama nya pelanggan.

4.1 Menampilkan Sebagian Isi Dataset

Untuk memberikan informasi tanpa keseluruhan dari dataset yang akan di analisis

Gambar 1: Menampilkan Sebagian Isi Dataset

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	..
0	5524	1957	Graduation	Single	58138.0	0	0	4/9/2012	58	635	.
1	2174	1954	Graduation	Single	46344.0	1	1	8/3/2014	38	11	.
2	4141	1965	Graduation	Together	71613.0	0	0	21/08/2013	26	426	.
3	6182	1984	Graduation	Together	26646.0	1	0	10/2/2014	26	11	.
4	5324	1981	PhD	Married	58293.0	1	0	19/01/2014	94	173	.

5 rows × 29 columns

4.1 *Cleaning Data*

Jika data sudah terkumpul maka dapat melakukan cleaning data bertujuan untuk memastikan dataset siap untuk di analisis lebih lanjut.

Gambar 2: Informasi Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   2240 non-null   int64  
 1   Year_Birth           2240 non-null   int64  
 2   Education            2240 non-null   object 
 3   Marital_Status       2240 non-null   object 
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64  
 6   Teenhome             2240 non-null   int64  
 7   Dt_Customer          2240 non-null   object 
 8   Recency              2240 non-null   int64  
 9   MntWines             2240 non-null   int64  
 10  MntFruits            2240 non-null   int64  
 11  MntMeatProducts      2240 non-null   int64  
 12  MntFishProducts      2240 non-null   int64  
 13  MntSweetProducts     2240 non-null   int64  
 14  MntGoldProds         2240 non-null   int64  
 15  NumDealsPurchases    2240 non-null   int64  
 16  NumWebPurchases      2240 non-null   int64  
 17  NumCatalogPurchases  2240 non-null   int64  
 18  NumStorePurchases    2240 non-null   int64  
 19  NumWebVisitsMonth    2240 non-null   int64  
 20  AcceptedCmp3         2240 non-null   int64  
 21  AcceptedCmp4         2240 non-null   int64  
 22  AcceptedCmp5         2240 non-null   int64  
 23  AcceptedCmp1         2240 non-null   int64  
 24  AcceptedCmp2         2240 non-null   int64  
 25  Complain             2240 non-null   int64  
 26  Z_CostContact        2240 non-null   int64  
 27  Z_Revenue            2240 non-null   int64  
 28  Response             2240 non-null   int64  
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

Dari output terdapat missing value pada kolom Income, pada kolom Dt_Customer yang menunjukkan tanggal seorang pelanggan bergabung dengan database belum diparsing sebagai DateTime. Terdapat beberapa fitul kategorikal dalam data frame. Oleh karen aitu perlu mengoonversi menjadi bentuk numerik.

Langkah awal yang harus dilakukan menghapus baris yang memiliki missing value pada kolom income. Setelah data missing value dibersihkan maka total data menjadi 2216 data. Selanjutnya membuat fitur baru dari Dt_Customer yang menunjukkan berapa lama seorang pelanggan telah terdaftar di database perusahaan.

Kita dapat mengeksplorasi nilai-nilai dalm fitur kategorikal untuk emndapatkan gambaran yang lebih jelas tenatang data

Gambar 3,4: Hasil Cleaning Data

```
Total categories in the feature Marital_Status:
 Marital_Status
Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: count, dtype: int64

Total categories in the feature Education:
 Education
Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: count, dtype: int64
```

	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000
mean	52247.251354	0.441787	0.505415	49.012635	305.091606	26.356047	166.995939	37.637635	27.028881	43.965253
std	25173.076661	0.536896	0.544181	28.948352	337.327920	39.793917	224.283273	54.752082	41.072046	51.815414
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	35303.000000	0.000000	0.000000	24.000000	24.000000	2.000000	16.000000	3.000000	1.000000	9.000000
50%	51381.500000	0.000000	0.000000	49.000000	174.500000	8.000000	68.000000	12.000000	8.000000	24.500000
75%	68522.000000	1.000000	1.000000	74.000000	505.000000	33.000000	232.250000	50.000000	33.000000	56.000000
max	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	262.000000	321.000000

4.2 Data Preprocessing

Data preprocessing dilakukan untuk melakukan oprasi clustering.

Gambar 6,7: Data Hasil Preprocessing Data

	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	...
0	-0.893586	0.287105	-0.822754	-0.929699	0.310353	0.977660	1.552041	1.690293	2.453472	1.483713	...
1	-0.893586	-0.260882	1.040021	0.908097	-0.380813	-0.872618	-0.637461	-0.718230	-0.651004	-0.634019	...
2	-0.893586	0.913196	-0.822754	-0.929699	-0.795514	0.357935	0.570540	-0.178542	1.339513	-0.147184	...
3	-0.893586	-1.176114	1.040021	-0.929699	-0.795514	-0.872618	-0.561961	-0.655787	-0.504911	-0.585335	...
4	0.571657	0.294307	1.040021	-0.929699	1.554453	-0.392257	0.419540	-0.218684	0.152508	-0.001133	...

5 rows × 23 columns

	count	mean	std	min	25%	50%	75%	max
col1	2212.0	4.497106e-17	2.878602	-5.978123	-2.539470	-0.781595	2.386380	7.452915
col2	2212.0	-1.927331e-17	1.709469	-4.194757	-1.323932	-0.173716	1.234923	6.168185
col3	2212.0	2.650080e-17	1.231685	-3.625184	-0.853556	-0.051292	0.863841	6.746845

4.3 Clustering

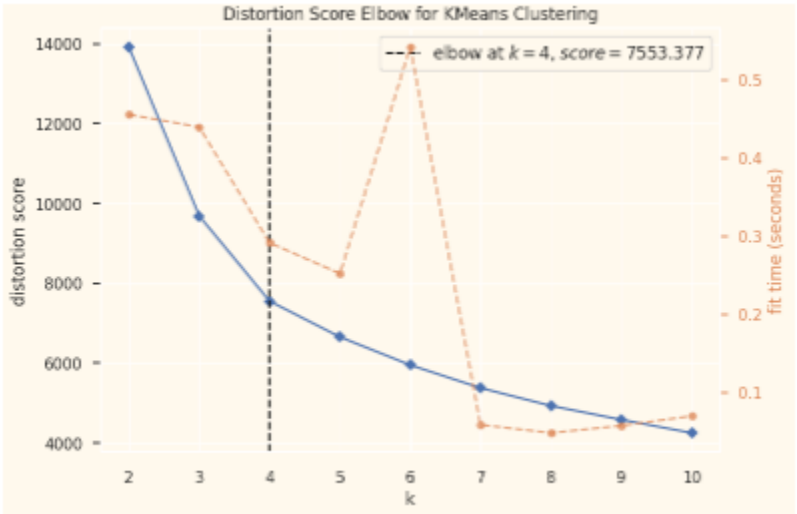
Langkah-langkah dalam clustering:

a. Metode elbow untuk menentukan jumlah cluster yang akan dibentuk

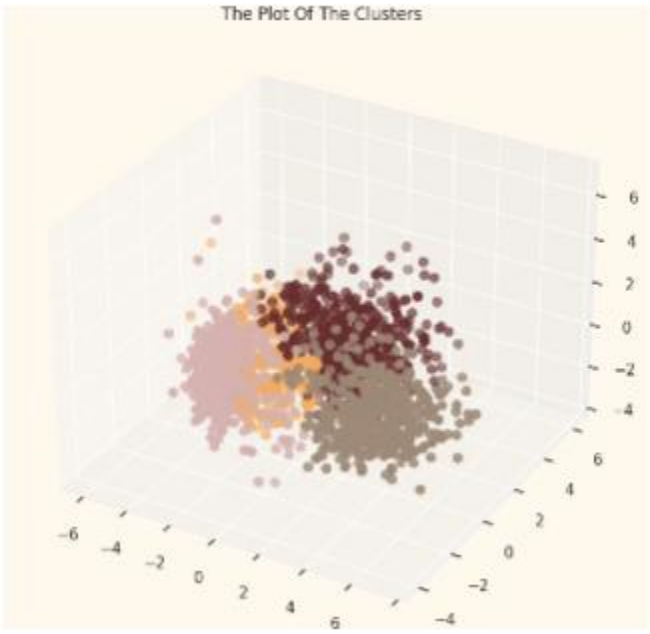
b. Melakukan clusterinng

c. Pemeriksaan custer yang terbentuk

Gambar 8,9: Grafik Clustering



<Axes: title={'center': 'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>



5. KESIMPULAN

Dari analisis yang sudah dilakukan maka kita mendapatkan 4 cluster segmentasi pelanggan yaitu.

Kluster 1: pelanggan denagn pendapatan tinggi dan pengeluaran tinggi, yang mayoritas memiliki anggota keluarga yang banyak

Kluster 2: pelanggan denagn pendapatan sedang dan pengeluaran moderat, umumnya terdiri dari keluarga kecil atau individu.

Kluster 3: pelanggan dengan pendapatan rendah dan pengeluaran rendah, sering kali indivisu atau kelaurga kecil denagn pengeluaran yang hemat.

Kluster 4: pelanggan denagn pola pengeluaran yang tidak knsisten, menunjukan pengeluara mereka tergantung pada promosi atau penawaran.

6. DAFTAR PUSTAKA

1. Karina, A., Mariza, K., (2019).. Penerapan Algoritma K-Means untuk Segmentasi Konsumen Menggunakan R. Jurnal Teknologi & Manajemen Informatika, 5(1).
2. Mustika, Yunita Ardilla, Abraham Manuhutu, Nazaruddin Ahmad, Imanuddin Hasbi, Guntoro, Melda Agnes Manuhutu, Mohamad Ridwan, Hozairi, Anindya Khrisna Wardhani, Syariful Alim, Ikhsan Romli, Yoga Religia, D Tri Octafian, Unggul Utan Sufandi , Iin Ernawati . (2021). Data Mining Dan Aplikasinya.
3. Rani Rotul Muhima, S.Si., M.T.,, Muchamad Kurniawan, S.Kom., M.Kom., Septiyawan Rosetya Wardhana, S.Kom., M.Kom., Anton Yudhana, S.T., M.T., Ph.D., Sunardi, S.T., M.T., Ph.D., Weny Mistarika Rahmawati, S.Kom., M.Kom.,M.Sc., Gusti Eka Yuliastuti, S.Kom., M.Kom. (2022). Kupas Tuntas Algoritma Clustering: Konsep, Perhitungan Manual, Dan Program.
4. Rojasqi Fadilla, Roni Andarsyah, Rolly Maulana Awangga, Roni Habibi. (2020). Data Analytics: Peningkatan Performa Algoritma Rekomendasi Collaborative Filtering Menggunakan K-Means Clustering.
5. Johan, O., (2013). Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University. Jurnal Ilmiah Teknik Industri, 12(1).
6. Beta, E. A., Indah, A., Adhistya, E. P. (2018). Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering, 2(1).
7. Ira, A., Reza, N. ., Lurinjani, A., Jerry, H., (2023). Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Pelanggan UHT Milk Greenfield. Cerdika: Jurnal Ilmiah Indonesia, 3(7).
8. Muwaddah, H., Yusniar, L., Zakarias, S., (2022). Analisis Pemasaran Bisnis dengan Data Science: Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering. DSI: Jurnal Data Science Indonesia, 1(2).