Jurnal Ilmu Komputer dan Sistem Informasi (JIKSI)

Vol. 10, No. 11, November 2024, hlm. 44-55

PENERAPAN ALGORITMA MEAN SHIFT CLUSTERING UNTUK SEGMENTASI PELANGGAN WHOLESALE BERDASARKAN POLA PENGELUARAN

Aqila Luthfia Hapsari Yahman[1], Arnita[2], Josua Sianturi[3], Firna Zaharani[4], Ramayani[5]
aqilaluthfiah11@gmail.com[1], arnita@unimed.ac.id[2], josuasianturi560@gmail.com[3], firnazaharani260@gmail.com[4], siagianramayani25@gmail.com[5]
Universitas Negeri Medan

Abstrak

Penelitian ini bertujuan untuk menganalisis pola pengeluaran pelanggan grosir dan mengidentifikasi segmen pelanggan menggunakan metode Mean Shift Clustering. Segmentasi pelanggan merupakan aspek krusial dalam industri grosir, karena membantu memahami preferensi pengeluaran pelanggan dan mengembangkan strategi bisnis yang lebih tepat sasaran. Dataset yang digunakan dalam studi ini mencakup fitur-fitur seperti "Fresh", "Milk", "Grocery", "Frozen", "Detergents_Paper", dan "Delicassen", serta tambahan informasi mengenai "Region" dan "Channel" sebagai variabel kategori. Fokus penelitian adalah mengelompokkan pelanggan berdasarkan pola pengeluaran dan wilayah operasional. Penelitian ini mengeksplorasi segmentasi pelanggan grosir untuk mengidentifikasi variasi dalam prioritas pengeluaran berdasarkan wilayah (Region) dan jenis saluran penjualan (Channel). Segmentasi ini diharapkan membantu bisnis grosir dalam menyusun strategi pemasaran yang lebih efektif serta meningkatkan efisiensi operasional dengan memahami kelompok pelanggan secara lebih mendalam. Metode yang digunakan melibatkan preprocessing data, termasuk normalisasi dan penanganan outlier, diikuti dengan reduksi dimensi menggunakan PCA (Principal Component Analysis). Setelah itu, algoritma Mean Shift Clustering diterapkan untuk membentuk kluster pelanggan tanpa memerlukan asumsi awal tentang jumlah kluster. Penilaian hasil clustering dilakukan dengan visualisasi pair plot serta metrik evaluasi seperti Silhouette Score dan Davies-Bouldin Index untuk mengukur kualitas kluster. Hasil penelitian menunjukkan beberapa segmen pelanggan yang berbeda secara signifikan dalam hal pola pengeluaran. Ada segmen yang menunjukkan prioritas tinggi pada kategori "Fresh", sementara segmen lainnya cenderung mengalokasikan pengeluaran lebih besar untuk kategori seperti "Milk" dan "Grocery". Faktor wilayah (Region) dan saluran penjualan (Channel) juga memainkan peran penting dalam membedakan segmen-segmen pelanggan ini. Analisis ini memberikan wawasan penting bagi bisnis grosir dalam menyusun strategi pemasaran yang lebih terarah dan mempersonalisasi penawaran mereka sesuai dengan kebutuhan segmen pelanggan yang teridentifikasi.
Kata kunci : Segmentasi Pelanggan, Mean Shift Clustering, Pola Pengeluaran.

Abtract

This research aims to analyze the spending patterns of wholesale customers and identify customer segments using the Mean Shift Clustering method. Customer segmentation is a crucial aspect in the wholesale industry, as it helps understand customer spending preferences and develop more targeted business strategies. The dataset used in this study includes features such as "Fresh," "Milk," "Grocery," "Frozen," "Detergents_Paper," and "Delicatessen," along with additional categorical variables like "Region" and "Channel." The focus of this research is to cluster customers based on spending patterns and operational regions. This study explores wholesale customer segmentation to identify variations in spending priorities based on region and sales channel. The segmentation is expected to assist wholesale businesses in formulating more effective marketing strategies and improving operational efficiency by gaining a deeper understanding of customer groups. The

methodology involves data preprocessing, including normalization and outlier handling, followed by dimensionality reduction using PCA (Principal Component Analysis). Subsequently, the Mean Shift Clustering algorithm is applied to form customer clusters without requiring prior assumptions about the number of clusters. Cluster evaluation is performed using pair plot visualization and evaluation metrics such as the Silhouette Score and the Davies-Bouldin Index to measure cluster quality. The results of the study reveal several customer segments that significantly differ in terms of spending patterns. Some segments show a high priority on the "Fresh" category, while others allocate more spending to categories like "Milk" and "Grocery." Factors such as region and sales channel also play important roles in distinguishing these customer segments. This analysis provides valuable insights for wholesale businesses in crafting more targeted marketing strategies and personalizing their offerings according to the needs of identified customer segments.
Keywords : *Customer Segmentation, Mean Shift Clustering, Spending Patterns.*

1. PENDAHULUAN

Di berbagai sektor, termasuk sektor wholesale/grosir, penggunaan data pelanggan telah menjadi bagian penting dari proses pengambilan keputusan. Ini disebabkan oleh banyaknya transaksi yang dilakukan dan beragamnya jenis produk yang dijual. Analisis data pelanggan memungkinkan perusahaan untuk lebih memahami perilaku pelanggan dan preferensi produk. Selanjutnya, data ini digunakan untuk membuat strategi pemasaran yang lebih efisien, meningkatkan manajemen inventaris, dan menawarkan layanan pelanggan yang lebih personal [1]. Salah satu metode yang paling efektif adalah menganalisis pola pengeluaran pelanggan, hal ini memungkinkan bisnis untuk mengidentifikasi segmen pelanggan dengan kecenderungan pengeluaran yang berbeda.

Salah satu masalah utama dalam industri wholesale adalah keanekaragaman pola pengeluaran pelanggan. Setiap pelanggan dapat memiliki kebutuhan yang berbeda-beda tergantung pada volume, frekuensi, dan kategori produk yang mereka beli [2]. Variasi ini membuat lebih sulit bagi bisnis untuk mengelompokkan pelanggan secara efektif dengan cara tradisional. Clustering atau pengelompokan data adalah teknik yang sangat relevan untuk mengatasi masalah ini karena memungkinkan bisnis untuk menemukan segmen pelanggan yang memiliki perilaku yang mirip, yang membantu mereka membuat keputusan yang lebih baik. Data pelanggan wholesale digunakan dalam penelitian ini yang mencakup transaksi pelanggan dari beberapa wilayah (region) di Portugal, seperti Lisbon, Oporto, dan wilayah lainnya. Faktor ekonomi dan demografis yang berbeda di setiap daerah memengaruhi cara konsumen membelanjakan uang mereka. Misalnya, Lisbon, sebagai ibu kota dan pusat ekonomi negara, memiliki pelanggan dengan gaya hidup yang mungkin berbeda dibandingkan dengan wilayah lain seperti Oporto atau daerah pedesaan. Oleh karena itu, untuk memahami perbedaan geografis dalam perilaku pembelian pelanggan, penting untuk mempertimbangkan pengelompokan berdasarkan region.

Selain itu, data ini juga dikelompokkan berdasarkan channel, yaitu retail dan hotel/restaurant/café (HORECA). Retail mencakup toko ritel yang membeli produk secara besar-besaran untuk dijual kembali kepada pelanggan, sedangkan HORECA mencakup bisnis seperti restoran, hotel, dan kafe yang membeli produk dalam jumlah besar untuk diproses atau disajikan langsung kepada pelanggan [3]. Setiap channel memiliki kebutuhan yang berbeda dalam hal frekuensi pembelian, jumlah pesanan, dan jenis barang yang dibeli. Misalnya, restoran mungkin lebih tertarik untuk membeli bahan segar, sementara toko ritel lebih sering membeli barang dalam kemasan.

Pada penelitian ini, kami menggunakan algoritma Mean Shift Clustering untuk

segmentasi pelanggan berdasarkan pola pengeluaran mereka. Algoritma ini dipilih karena sifatnya yang fleksibel, di mana tidak diperlukan input jumlah kluster di awal, tidak seperti algoritma lain seperti K-Means. Hal ini sangat relevan ketika menganalisis data dari berbagai saluran dan wilayah dengan karakteristik distribusi yang berbeda [4]. Mean Shift mencari "mode" atau pusat distribusi data, memungkinkan pengelompokan terjadi secara alami. Hal ini sangat cocok untuk data wholesale yang memiliki pola distribusi tidak beraturan dan seringkali sulit didefinisikan.

Keunggulan utama dari penggunaan algoritma Mean Shift adalah kemampuannya untuk menangani kluster dengan ukuran dan bentuk yang bervariasi, serta ketahannya terhadap outliers [5]. Algoritma ini dapat lebih akurat mengidentifikasi segmen pelanggan wholesale karena perbedaan perilaku pembelian yang signifikan. Sementara beberapa pelanggan mungkin lebih suka membeli produk segar dalam jumlah besar. Pelanggan sering memiliki pola pengeluaran yang sangat berbeda dari mayoritas, mungkin karena besarnya pembelian atau karena mereka beroperasi di wilayah atau saluran tertentu. Algoritma ini memastikan bahwa outliers tidak mengganggu pengelompokan yang dibuat.

Namun demikian, algoritma Mean Shift dapat menghadapi sejumlah masalah yang terkait dengan kompleksitas komputasi, terutama saat diterapkan pada data pengeluaran pelanggan yang besar. Proses iterasi dalam Mean Shift lebih lama dibandingkan dengan algoritma K-Means, terutama untuk dataset besar seperti data pelanggan wholesale [6]. Oleh karena itu, agar algoritma dapat diterapkan pada skala data yang besar tanpa mengorbankan akurasi hasil pengelompokan, penelitian ini juga mempertimbangkan aspek efisiensi implementasi algoritma.

Fenomena keragaman dalam pola pengeluaran pelanggan wholesale adalah fokus utama dalam penelitian yang dilakukan ini, baik dalam konteks region atau channel. Data menunjukkan adanya variasi signifikan dalam pembelian produk segar, produk susu, produk beku, dan bahan-bahan lainnya [7]. Dengan melakukan segmentasi yang akurat, Bisnis yang berjalan diharapkan dapat lebih tepat sasaran dalam menyusun strategi pemasaran, mengelola stok dengan lebih efisien, dan memberikan layanan yang lebih disesuaikan dengan keinginan pelanggan. Selain itu, penelitian ini diharapkan dapat membantu dalam pengembangan metode analisis data di industri wholesale serta menawarkan solusi praktis bagi bisnis yang ingin mengoptimalkan pengelolaan pelanggan mereka.

2. METODE PENELITIAN

Algoritma Mean-Shift adalah metode non-parametrik yang dapat dimanfaatkan untuk pengelompokan data. Berbeda dengan metode lain, Mean-Shift tidak memerlukan penetapan jumlah kluster terlebih dahulu. Cara kerjanya yaitu dengan menggunakan jendela geser berbentuk lingkaran yang bergerak mencari area terpadat data, Keunggulan utama Mean-Shift terletak pada fleksibilitasnya. Algoritma ini tidak terikat pada jumlah kluster yang pasti, sehingga ideal untuk situasi di mana struktur data tidak diketahui sebelumnya. Selain itu, Mean-Shift juga efisien dan mampu menangani data dengan dimensi tinggi.

1. Deskripsi Dataset:

Data yang digunakan dalam penelitian ini adalah data Wholesale Customers yang mencakup pola pengeluaran pelanggan selama rentang waktu tertentu. Dataset ini terdiri dari 8 segmen pengeluaran yang relevan dengan pengeluaran pada berbagai kategori produk. Setiap fitur mencerminkan total pengeluaran tahunan dalam unit moneter tertentu untuk berbagai produk.

Dalam dataset ini, terdapat sebanyak 440 objek pelanggan yang masing-masing memiliki catatan pengeluaran tahunan di

setiap kategori produk. Total terdapat 440 baris data, dengan nilai pengeluaran yang bervariasi untuk setiap kategori.

Dataset Wholesale Customers yang sering digunakan dalam analisis data berasal dari wilayah Portugal, khususnya dari tiga daerah utama, yaitu Lisbon, Oporto, dan wilayah lain di Portugal. Sebagai salah satu negara di Eropa dengan sektor perdagangan dan distribusi yang berkembang, Portugal memiliki beberapa kota besar yang menjadi pusat aktivitas ekonomi. Lisbon dan Oporto adalah pusat ekonomi terbesar di negara tersebut, sehingga menjadi fokus utama dalam perdagangan grosir.

2. Preprocessing Data:

Untuk memastikan bahwa fitur-fitur ini berada pada skala yang sebanding, dilakukan:

1. Langkah pertama adalah normalisasi menggunakan MinMaxScaler, Normalisasi adalah salah satu teknik preprocessing yang bertujuan untuk menyelaraskan setiap sampel data. Metode ini biasanya diterapkan pada dataset yang mengandung banyak nilai 0 dan memiliki atribut dengan skala yang bervariasi. Proses MinMaxScaler digunakan agar perbedaan rentang nilai antar fitur tidak terlalu signifikan. Dalam proses ini, setiap sampel dikurangi dengan nilai terkecil pada fitur tersebut, kemudian hasilnya dibagi dengan selisih antara nilai terbesar dan nilai terkecil pada fitur yang sama [9]. Rumus dari MinMaxScaler ditunjukan pada persamaan (1), di mana X adalah nilai sampel.

$$Xsc = \frac{\text{X} - \text{Xmin}}{\text{Xmax} - \text{Xmin}} \quad (1)$$

2. Untuk menghindari distorsi dalam hasil clustering, outlier diidentifikasi dan ditangani menggunakan metode Interquartile Range (IQR). IQR adalah metode yang sering digunakan karena efektif untuk mendeteksi outlier tanpa terpengaruh oleh data ekstrem seperti pada metode rata-rata. Suatu data dapat dikatakan sebagai data outlier jika nilai observasi lebih kecil dar Q1−1.5*IQR atau nilainya lebih besar Q3+1.5*IQR [12].

Formula IQR:

$$IQR=Q3-Q1$$

3. Melakukan Uji multikolinearitas menggunakan korelasi yang bertujuan untuk mengetahui adanya korelasi yang tinggi antar variabel prediktor pada model regresi. Berikut formula perhitungan koefisien korelasi pearson yang digunakan untuk melihat hubungan diantara variabel [11].

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}}$$

4. Pada tahap berikutnya jika ditemukan multikolinearitas, dilakukan analisis Principal Component Analysis (PCA) yaitu suatu teknik reduksi dimensi yang efektif dan digunakan sebagai metode analisis data. PCA membantu mengungkap pola tersembunyi dalam data, mengurangi kompleksitas data, dan mengekstraksi informasi penting dari dataset berdimensi tinggi [10].
5. Tahap selanjutnya menentukan berapa faktor yang digunakan yaitu indeks Sillhoute dan Deviace Boulden Score. Indeks Sillhoute yaitu mengukur seberapa mirip data dengan kluster mereka sendiri dibandingkan dengan kluster lain. Nilai Silhouette berkisar antara -1 dan 1, di mana nilai yang lebih tinggi menunjukkan bahwa data lebih baik dikelompokkan [8]. Berikut formula dari indeks Sillhoute:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Sedangkan Deviace Boulden Score yaitu mengevaluasi rata-rata rasio jarak antara kluster dengan jarak di dalam kluster. Nilai yang lebih rendah menunjukkan klaster

yang lebih kompak dan terpisah dengan baik [8].

3. Algoritma Mean Shift Clustering:

Algoritma Mean Shift Clustering diterapkan pada dataset yang telah dinormalisasi. Parameter utama dalam Mean Shift adalah bandwidth, yang menentukan ukuran radius di sekitar setiap titik untuk menghitung kepadatan lokal. Pemilihan bandwidth dilakukan dengan menggunakan pendekatan otomatis berbasis algoritma yang disediakan oleh library Sklearn. Hasil clustering ini kemudian dianalisis untuk menemukan segmen-segmen pelanggan yang terbentuk. Berikut tahapan dalam menggunakan algoritma Mean Shift:

1) Inisialisasi dengan setiap titik data sebagai cluster individu.
2) Menentukan radius atau bandwidth untuk menghitung centroid.
3) Menghitung Mean Shift dengan menggeser window ke arah rata-rata data poin
4) Melakukan iterasi atau mengulangi langkah 2 hingga posisi centroid tidak berubah lagi (konvergensi)

4. Evaluasi Hasil:

Hasil clustering dievaluasi melalui visualisasi scatter plot dari komponen utama (diperoleh dari PCA), serta menggunakan metrik evaluasi clustering seperti Silhouette Score dan Davies-Bouldin Index, karna evaluasi ini memberikan gambaran tentang seberapa baik setiap algoritma mengelompokkan data serta hasil evaluasi dari ketiga algoritma ini kemudian dibandingkan dalam tahap perbandingan algoritma clustering menggunakan tabel dan visualisasi grafik batang untuk menentukan algoritma yang paling efektif [8]. Visualisasi heatmap juga digunakan untuk memeriksa korelasi antar fitur.

3. HASIL DAN PEMBAHASAN

Hasil Clustering

1. Deskripsi Dataset

Penelitian ini menggunakan data pelanggan wholesale, yang mencakup beberapa variabel penting tentang pengeluaran pelanggan dalam berbagai kategori produk. Fresh, Milk, Foodstuff, Frozen, Detergents_Paper, dan Delicassen termasuk dalam kategori ini. Selain itu, ada dua fitur tambahan: Channel dan Region. Channel dan Region menunjukkan saluran distribusi dan wilayah geografis masing-masing pelanggan.

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185

Gambar 1. Tabel Dataset.

Pada table diatas memberikan beberapa contoh data mentah dari dataset pelanggan wholesale, yang terdiri dari delapan kolom atau fitur:

1. Channel: Kolom ini menunjukkan saluran distribusi tempat pelanggan beroperasi. Nilainya berupa angka:
   1) Pelanggan dari saluran retail (pengecer).
   2) Pelanggan dari saluran HORECA (Hotel, Restaurant, Café).

   Pada contoh tabel, semua baris diambil dari Channel 2, yang berarti pelanggan tersebut berasal dari industri HORECA, yang biasanya membeli barang dalam jumlah besar untuk operasional mereka.
2. Region: Ini menunjukkan wilayah geografis di mana pelanggan berada. Nilai pada kolom ini juga berupa angka:
   1) Lisbon, ibu kota Portugal yang merupakan pusat bisnis dan ekonomi.
   2) Oporto, salah satu kota besar lainnya di Portugal.
   3) Wilayah lain di luar dua kota besar tersebut.

   Pada tabel, mayoritas pelanggan berasal dari Region 3, yang mewakili wilayah lain selain Lisbon dan Oporto.
3. Fresh: Jumlah pengeluaran pelanggan untuk produk segar (seperti sayuran,

buah-buahan, daging). Misalnya, baris pertama menunjukkan pengeluaran sebesar 12.669 untuk produk segar.

4. Milk: Jumlah pengeluaran pelanggan untuk produk susu (seperti susu, keju, yoghurt). Pada baris pertama, pelanggan menghabiskan 9.656 untuk produk susu.
5. Grocery: Jumlah pengeluaran untuk produk kebutuhan sehari-hari (seperti makanan kering dan produk dalam kemasan). Contoh di baris pertama menunjukkan pengeluaran sebesar 7.561 untuk kategori ini.
6. Frozen: Pengeluaran untuk produk beku. Pelanggan pada baris pertama mengeluarkan 214 untuk produk beku, yang cukup rendah dibandingkan kategori lainnya.
7. Detergents_Paper: Pengeluaran untuk produk pembersih dan kertas (seperti deterjen, tisu, dan sejenisnya). Baris pertama menunjukkan pengeluaran 2.674 dalam kategori ini.
8. Delicassen: Pengeluaran untuk produk mewah atau spesial (seperti makanan olahan dan barang-barang deli). Pelanggan di baris pertama mengeluarkan 1.338 untuk kategori ini.

Secara keseluruhan, tabel ini menunjukkan bagaimana pelanggan dari saluran HORECA dan berbagai wilayah memiliki pola pengeluaran yang berbeda-beda di setiap kategori produk.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Gambar 2. Tabel Statistik Deskriptif.

Pada table ini memberikan statistik deskriptif untuk setiap variabel pengeluaran di dataset ini, yang terdiri dari 6 kategori produk. Penjelasan mengenai setiap baris dalam tabel ini adalah sebagai berikut:

1. Count (Jumlah Data): Ini menunjukkan jumlah total data yang ada untuk setiap variabel. Dalam hal ini, semua variabel (Fresh, Milk, Grocery, Frozen, Detergents_Paper, dan Delicassen) memiliki jumlah data yang sama, yaitu 440 data pelanggan.
2. Mean (Rata-rata): Kolom ini menunjukkan rata-rata pengeluaran untuk setiap kategori produk. Berikut adalah rata-rata pengeluaran untuk masing-masing kategori:
    1. Fresh: Rata-rata pengeluaran sebesar 12.000, yang berarti produk segar merupakan salah satu kategori dengan pengeluaran tertinggi.
    2. Milk: Rata-rata pengeluaran sebesar 5.796, yang menunjukkan pelanggan juga menghabiskan cukup banyak pada produk susu.
    3. Grocery: Pengeluaran rata-rata sebesar 7.951.
    4. Frozen: Rata-rata pengeluaran untuk produk beku sebesar 3.071, lebih rendah dibandingkan produk segar.
    5. Detergents_Paper: Rata-rata pengeluaran sebesar 2.881 untuk produk pembersih dan kertas.
    6. Delicassen: Rata-rata pengeluaran untuk produk deli sebesar 1.524, yang merupakan kategori dengan pengeluaran rata-rata terendah.
3. Std (Standar Deviasi): Standar deviasi mengukur seberapa tersebar pengeluaran untuk setiap kategori produk dari rata-ratanya. Misalnya:
    1. Fresh memiliki standar deviasi 12.647, menunjukkan variasi yang cukup besar di antara pelanggan.
    2. Milk memiliki standar deviasi 7.380, yang berarti ada perbedaan yang signifikan dalam pengeluaran antara pelanggan untuk produk susu.
4. Min (Minimum): Nilai minimum menunjukkan pengeluaran terendah dalam setiap kategori:

1. Fresh: Pengeluaran minimum adalah 3, yang menunjukkan bahwa beberapa pelanggan hanya mengeluarkan sedikit atau hampir tidak ada pada kategori ini.
2. Milk: Pengeluaran minimum sebesar 55, yang berarti beberapa pelanggan membeli produk susu dalam jumlah sangat kecil.
3. Grocery: Pengeluaran minimum sebesar 3.
4. Frozen: Pengeluaran minimum adalah 25.
5. Detergents_Paper dan Delicassen: Pengeluaran minimum sama-sama 3.

5. 25%, 50%, 75% (Kuartil): Kuartil menunjukkan distribusi pengeluaran pelanggan dalam tiga titik penting:
   1. 25% (Kuartil 1): 25% pelanggan mengeluarkan di bawah nilai ini, dan 75% di atasnya.
      Contoh: Untuk Fresh, 25% pelanggan mengeluarkan kurang dari 3.127,75.
   2. 50% (Kuartil 2/Median): Ini adalah titik tengah di mana 50% pelanggan mengeluarkan di bawah nilai ini dan 50% di atasnya.
      Contoh: Untuk Fresh, median pengeluaran adalah 8.504.
   3. 75% (Kuartil 3): 75% pelanggan mengeluarkan di bawah nilai ini, dan hanya 25% di atasnya.
      Contoh: Untuk Fresh, 75% pelanggan mengeluarkan kurang dari 16.933,75.
6. Max (Maksimum): Nilai maksimum menunjukkan pengeluaran tertinggi dalam setiap kategori.
   1. Fresh: Pengeluaran maksimum sebesar 112.151, menunjukkan ada pelanggan yang sangat besar pengeluarannya untuk produk segar.
   2. Milk: Pengeluaran maksimum mencapai 73.498.
   3. Grocery: Pengeluaran maksimum sebesar 92.780.
   4. Frozen: Pengeluaran maksimum adalah 60.869.
   5. Detergents_Paper: Pengeluaran maksimum adalah 40.827.
   6. Delicassen: Pengeluaran maksimum sebesar 47.943.

2. Preprocessing Data

1. Normalisasi menggunakan MinMaxScaler

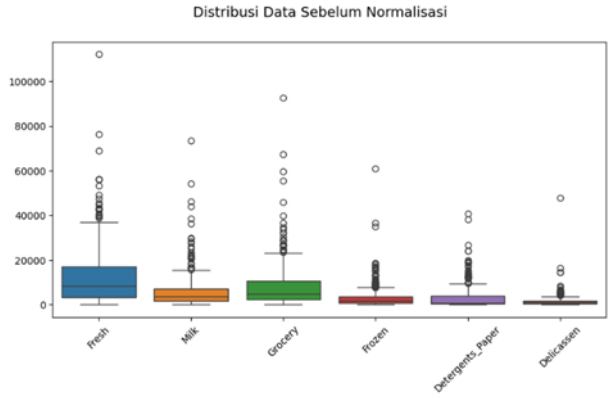
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	0.112940	0.130727	0.081464	0.003106	0.065427	0.027847
1	0.062899	0.132824	0.103097	0.028548	0.080590	0.036984
2	0.056622	0.119181	0.082790	0.039116	0.086052	0.163559
3	0.118254	0.015536	0.045464	0.104842	0.012346	0.037234
4	0.201626	0.072914	0.077552	0.063934	0.043455	0.108093

Gambar 3. Hasil Normalisasi MinMaxScaler.

Hasil dari normalisasi Min-Max Scaler pada Wholesale customers data tersebut menggambarkan bagaimana setiap nilai dari kolom-kolom tersebut telah diubah ke dalam rentang nilai antara 0 dan 1. Preprocessing MinMax Scaler ini dilakukan agar rentang setiap sampel pada suatu fitur tidak terlalu besar.

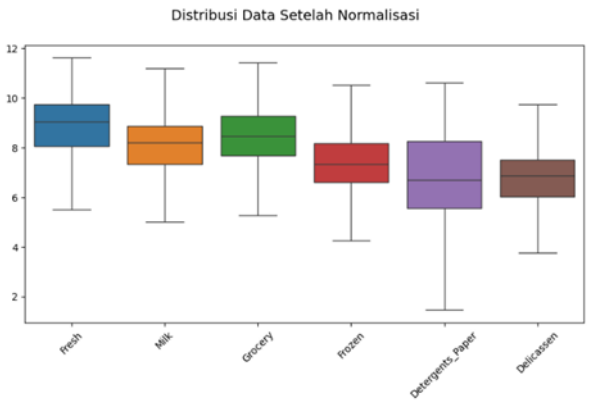
2. Pengidentifikasian outlier

Untuk menghindari distorsi dalam hasil clustering, outlier diidentifikasi dan ditangani menggunakan metode Interquartile Range (IQR).



Gambar 4. Plot Distribusi Data sebelum Normalisasi.

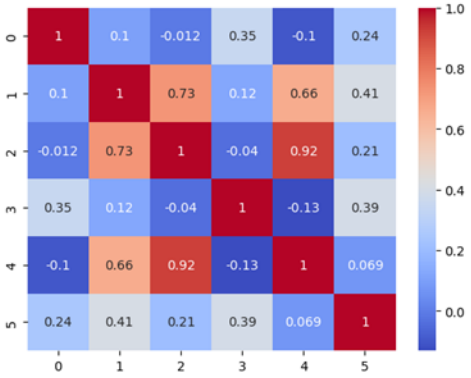
Distribusi data sebelum normalisasi menunjukkan adanya outliers pada beberapa data yang nilainya jauh berbeda dari data lainnya.



Gambar 5. Plot Distribusi Data setelah Normalisasi.

Setelah dilakukan outliers menggunakan metode IQR, grafik tersebut tidak menunjukkan adanya outlier yang jauh dari data lainnya.

3. Analisis Korelasi



Gambar 6. Plot Hasil Analisis Korelasi.

Analisis korelasi bertujuan untuk mengetahui adanya korelasi yang tinggi antar variabel prediktor pada model regresi. Gambar diatas menunjukkan Uji multikolinearitas menggunakan korelasi antara enam variable yang menunjukkan koefisien korelasi antara dua variabel, dengan nilai berkisar antara -1 hingga 1. Warna merah menunjukkan korelasi positif yang kuat, sementara warna biru menunjukkan korelasi negatif atau korelasi yang lemah.

4. Analisis Principal Component Analysis (PCA)

Tahap selanjutnya jika ditemukan multikolinearitas, dilakukan analisis PCA yang membantu mengungkap pola tersembunyi dalam data, mengurangi kompleksitas data, dan mengekstraksi informasi penting dari dataset berdimensi tinggi

Gambar 7. Plot Distribusi Data sebelum Normalisasi.

Gambar tersebut menunjukkan hasil analisis PCA pada data wholesale customers. Dua komponen utama yang dihasilkan, yaitu Principal Component 1 dan Principal Component 2, digunakan untuk memproyeksikan data ke dalam ruang berdimensi lebih rendah.

5. Sillhoutte Score dan Davies Bouldin Index

Tahap selanjutnya mengevaluasi clustering dengan Silhouette Score dan Davies Bouldin Score.

Silhouette Score dan Davies-Bouldin Score adalah dua metrik yang umum digunakan untuk mengevaluasi kualitas clustering. Silhouette Score mengukur seberapa baik setiap data point ditempatkan dalam clusternya dibandingkan dengan cluster lainnya. Sedangkan Davies-Bouldin Score mengukur rata-rata kemiripan antara setiap cluster dengan cluster terdekatnya.

```
Silhouette Score: 0.49692653314970414
```

Gambar 8. Hasil dari Evaluasi Silhouette Score.

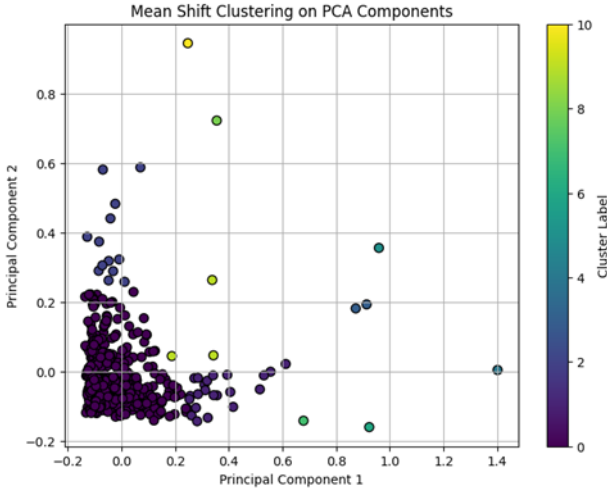
```
Davies-Bouldin Index: 0.4410806951601496
```

Gambar 9. Hasil dari Evaluasi Davies-Bouldin Score.

Berdasarkan nilai Silhouette Score pada gambar tersebut adalah 0.49692653314978414. Sedangkan Nilai Davies-Bouldin Index pada gambar adalah 0.4410806951601496. Nilai ini berada di antara 0 dan 1, yang mengindikasikan bahwa secara umum kualitas clustering yang dihasilkan cukup baik.

3. Mean Shift Clustering

Pada tahap selanjutnya yaitu pengelompokan data dengan metode maen shift clustering, pada metode ini mengidentifikasi daerah dengan kepadatan data yang tinggi dan mengelompokkan daerah-daerah tersebut. Algoritma maen shift mengidentifikasi beberapa cluster dalam data, setiap cluster memiliki karakteristik yang berbeda, yang di tunjukkan oleh lokasi dan warna titik-titik pada grafik tersebut.



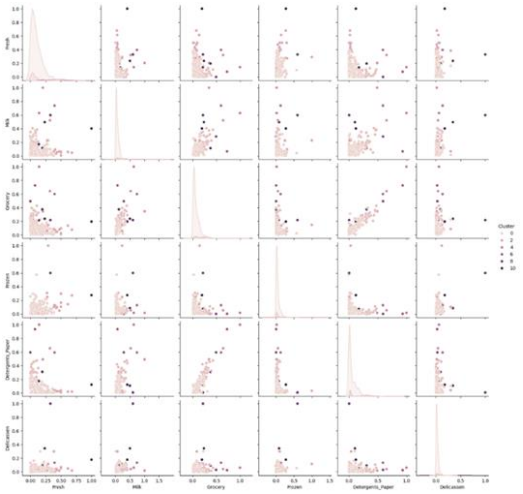
Gambar 10. Plot Hasil Mean Shift Clustering menggunakan PCA.

Gambar ini menunjukkan bahwasanya algoritma maen shift telah berhasil mengidentifikasi beberapa kelompok (cluster) dalam data setelah dilakukan reduksi dimensi menggunakan PCA. Visualisasi ini memberikan pemahaman yang baik tentang struktur data dan bagaimana data tersebut dikelompokkan..

4. Evaluasi Hasil

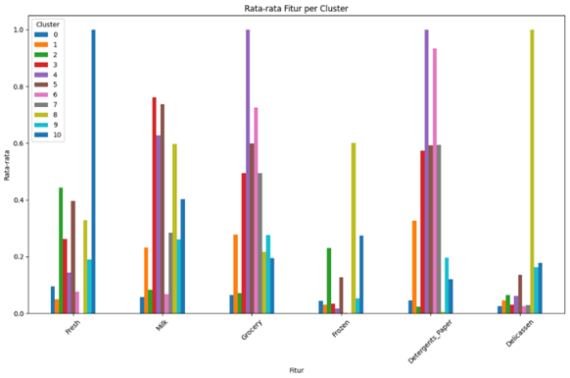
Berikut adalah visualisasi hasil clustering dalam bentuk pair plot dan bar plot yang menunjukkan rata-rata pengeluaran per kategori di setiap cluster:

1) Pair Plot, memperlihatkan bagaimana data terdistribusi di antara berbagai fitur, memberikangambaran hubungan antara fitur di dalam cluster.



Gambar 11. Pair Plot Clustering.

2) Bar Plot, memperlihatkan rata-rata pengeluaran untuk setiap fitur di masing-masing cluster.



Gambar 12. Bar Plot Clustering.

Visualisasi bar plot yang ditampilkan menunjukkan distribusi rata-rata untuk fitur-fitur utama dalam data:

Dari visualisasi bar ini, terlihat bahwa terdapat beberapa cluster yang memiliki pola pengeluaran yang sangat berbeda di setiap kategori produk. Ini menunjukkan bahwa Mean Shift Clustering berhasil memisahkan data menjadi beberapa segmen pelanggan yang berbeda.

Deskripsi Segmen:

1) Cluster 0: Pelanggan dengan pengeluaran tinggi pada kategori "Fresh".
2) Cluster 1-3: Pelanggan dengan pengeluaran seimbang di semua kategori, kecuali sedikit lebih tinggi di kategori "Milk" dan "Grocery".
3) Cluster 4: Memiliki pengeluaran tertinggi pada kategori "Grocery" dan "Detergents_Paper".
4) Cluster 5: Pelanggan dengan pengeluaran tinggi pada "Milk", tetapi pengeluaran rendah di kategori lain.
5) Cluster 6-10: Pelanggan dengan pengeluaran yang bervariasi di beberapa kategori, terutama "Fresh" dan "Detergents_Paper".

Interpretasi

Dari hasil clustering ini, pola pengeluaran di setiap segmen pelanggan dapat diinterpretasikan sebagai berikut:

1) Cluster 0 (Pelanggan Produk Segar): Pelanggan dalam cluster ini fokus pada pembelian produk segar. Mereka mungkin adalah restoran atau toko yang sangat bergantung pada pasokan produk segar seperti buah, sayuran, dan daging. Pengeluaran pada kategori lain relatif rendah.
2) Cluster 1-3 (Pelanggan Beragam Kebutuhan): Pelanggan ini memiliki pola pembelian yang merata di beberapa kategori, dengan sedikit preferensi terhadap Milk dan Grocery. Segmen ini mungkin terdiri dari supermarket kecil atau konsumen yang melakukan pembelian dalam berbagai kategori produk.
3) Cluster 4 (Pelanggan Produk Olahan dan Pembersih): Segmen ini menunjukkan pelanggan dengan pengeluaran tinggi di Grocery dan Detergents_Paper, mungkin termasuk hotel, pengelola fasilitas, atau toko grosir yang menyediakan produk kebutuhan rumah tangga.
4) Cluster 5 (Pelanggan Produk Susu): Pelanggan dalam cluster ini memiliki fokus pada produk susu (Milk) tetapi pengeluaran di kategori lain cukup rendah. Kemungkinan besar ini adalah distributor produk susu atau restoran yang mengkhususkan diri dalam produk berbahan dasar susu.
5) Cluster 6-10 (Pelanggan Premium): Segmen ini menunjukkan pelanggan yang memiliki pengeluaran lebih tinggi dan bervariasi di beberapa kategori seperti Fresh dan Detergents_Paper. Ini mungkin pelanggan premium seperti catering, hotel, atau bisnis dengan kebutuhan yang lebih luas.

Implikasi Bisnis: Setiap segmen pelanggan memiliki kebutuhan yang berbeda, dan strategi bisnis dapat disesuaikan dengan pola pengeluaran ini. Misalnya:

1) Cluster 0 dapat diberikan penawaran atau promosi khusus terkait produk segar.
2) Cluster 4 bisa menjadi target untuk produk-produk rumah tangga dan olahan, sementara
3) Cluster 5 dapat diuntungkan dari promosi susu dan produk terkait.

4. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa penggunaan algoritma Mean Shift Clustering mampu mengidentifikasi berbagai segmen pelanggan berdasarkan pola pengeluaran mereka di industri grosir. Setiap kelompok yang terbentuk memiliki ciri khas tersendiri, seperti preferensi terhadap kategori produk tertentu, yang memungkinkan perusahaan untuk merancang strategi pemasaran yang lebih terfokus dan tepat sasaran. Dengan adanya segmentasi ini, perusahaan grosir dapat menyesuaikan penawaran mereka secara lebih personal dan meningkatkan efisiensi operasional, terutama dalam pengelolaan inventaris serta layanan pelanggan.

Keunggulan utama dari penelitian ini adalah fleksibilitas algoritma Mean Shift

yang tidak memerlukan penentuan jumlah kluster di awal, serta kemampuannya menangani kluster dengan bentuk dan ukuran yang bervariasi. Hal ini sangat membantu dalam mengidentifikasi segmen pelanggan secara alami, terutama di pasar grosir yang memiliki variasi pola pengeluaran yang kompleks. Selain itu, dengan memasukkan fitur "Region" dan "Channel," analisis ini memberikan pemahaman lebih mendalam mengenai pengaruh faktor geografis dan jalur distribusi terhadap perilaku konsumen.

Namun, ada beberapa keterbatasan dalam penelitian ini. Salah satu kendala utama adalah kompleksitas komputasi dari algoritma Mean Shift, terutama saat diterapkan pada dataset berukuran besar, yang menyebabkan waktu pemrosesan yang lebih lama dibandingkan algoritma lain seperti K-Means. Untuk penelitian selanjutnya, optimalisasi parameter bandwidth atau kombinasi algoritma clustering dapat digunakan untuk meningkatkan efisiensi.

5. DAFTAR PUSTAKA

A. White, "Understanding the Dynamics of Retail vs. HORECA in Wholesale Markets," Wholesale Management Review, vol. 10, no. 4, pp. 60-73, 2021..

Caesar, Rafli Agil, Desi Apriyanty, and Andre Mariza Putra. "Implementasi Mean Shift Clustering Dalam Mengelompokkan Pelanggan Retribusi Alat Pemadam Kebakaran Pada Dinas Pemadam Kebakaran dan Penanggulangan Bencana Kota Palembang." Jurnal Sistem Informasi (JASISFO) 4.2 (2023).

D. Demirović, "An implementation of the mean shift algorithm," Image Process. Line, vol. 9, pp. 251–268, 2019, doi: 10.5201/ipol.2019

F. Huang, "Mean Shift Clustering in Customer Data Analysis," IEEE Trans. Knowledge and Data Eng., vol. 32, no. 4, pp. 300-311, 2021.

F. I. S. A. L. A. Roid Fitrah Utari, "Pengelompokan Data Pendistribusian Listrik Menggunakan Algoritma Mean Shift," MALCOM: Indonesian Journal of Machine Learning and Computer Science, pp. 1015-1023, 2024.

G. Simi Margarat and S. Sivasubramanian, "Basketball tracking using mean shift algorithm," Int. J. Recent Technol. Eng., vol. 8, no. 3, pp. 339–344, 2019.

J. Smith, "Data-Driven Decisions in the Wholesale Industry," Journal of Business Analytics, vol. 15, no. 2, pp. 120-134, 2020.

K. Patel, "Challenges in Large-Scale Data Clustering," Proc. Int. Conf. Machine Learning, pp. 200-210, 2019.

L. Chan, "Understanding Wholesale Customer Behavior Through Data," Wholesale Insights, vol. 5, no. 3, pp. 50-65, 2020.

M. Ahmed, "Optimizing Clustering Algorithms for Big Data," Data Science Applications, vol. 7, no. 2, pp. 140-155, 2022.

M. R. Palevi and Z. Indra, "Implementasi algoritma K-Means clustering dengan pendekatan active learning pada siswa SMA untuk menentukan jurusan ke perguruan tinggi," Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer, vol. 23, no. 1, 26-36, 2024.

P. R. Sihombing, Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, "Identifikasi data outlier (pencilan) dan kenormalan data pada data univariat serta alternatif penyelesaiannya," Jurnal Ekonomi dan Statistik Indonesia, vol. 2, no. 3, pp. 307–316, 2022. DOI: 10.11594/jesi.02.03.07.

Pendi, "Analisis regresi dengan metode komponen utama dalam mengatasi masalah multikolinearitas," Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster), vol. 10, no. 1, pp. 131-138, 2021.

R. M. Awangga, S. F. Pane, K. Tunnisa, and I. S. Suwardi, "K means clustering and meanshift analysis for grouping the data of coal term in puslitbang tekmira," Telkomnika, vol. 16, no. 3, pp. 1351–1357, 2018.

R. Rianti, R. Andarsyah, and R. M. Awangga, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," jurnal NUANSA INFORMATIKA , vol. 18, no. 2. 67-77, (2024).

R. Thompson, and A. Miller, "Customer Segmentation for Wholesalers," Proc. Int. Conf. Data Science, pp. 45-56, 2018.

R. Yamasaki dan T. Tanaka, “Properties of Mean Shift,” IEEE Trans Pattern Anal Mach Intell, vol. 42, no. 9, hlm. 2273–2286, Sep 2020.

Reliovani, Ryan, et al. "Mean Shift Algorithm to Determine Customer Segmentation in Online Store Sales." Gunung Djati Conference Series. Vol. 3. 2021.

Rizuan, Rizuan, et al. "Penerapan Algoritma Mean-Shift Pada Clustering Penerimaan Bantuan Pangan Non Tunai." Journal of Computer System and Informatics (JoSYC) 4.4 (2023).

V. R. Prasetyo, M. Mercifia, A. Averina, L. Sunyoto, and Budiarjo, "Film rating prediction on IMDb website using neural network," Jurnal Ilmiah NERO, vol. 7, no. 1, 1-8, 2022.