JURNAL MATEMATIKA SAINS SAHOMBU

E-ISSN 25782-78

SAHOMBU

PENERAPAN ALGORITMA DBSCAN UNTUK ANALISIS DEMOGRAFIS dan PENGELUARAN PELANGGAN MALL

Yusfi Syawali [1]*, Arnita [2], Muhammad Haikal Hafiz Rangkuti [3], Kaka Aprianda Mayadi [4], Kaka Silvi Pratiwi [5], Muhammad Fahrur Razi [6]

[1,2,3,4,5,6] Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Negeri Medan, Indonesia

Abstrak.

Dengan menganalisis dataset yang terdiri dari 200 entri, penelitian ini mengeksplorasi demografi pelanggan di pusat perbelanjaan. Variabel yang dievaluasi mencakup gender, usia, pendapatan tahunan (k$), dan skor pengeluaran (1 hingga 100). Hasil analisis deskriptif menunjukkan bahwa usia rata-rata pelanggan adalah 38,85 tahun, dengan pendapatan tahunan rata-rata sebesar 60,56 ribu dolar dan skor pengeluaran rata-rata 50,2, serta memastikan tidak ada nilai hilang dalam dataset. Analisis distribusi mengungkapkan perbedaan signifikan antara pria dan wanita dalam hal pendapatan dan pengeluaran. Untuk segmentasi pelanggan, algoritma DBSCAN menghasilkan lima cluster utama dan satu cluster outlier. Kombinasi parameter optimal untuk model ditemukan pada eps = 12,5 dan min_samples = 4, dengan skor silhouette tertinggi. Meskipun terdapat korelasi negatif kecil antara usia dan skor pengeluaran, usia tidak berfungsi sebagai prediktor yang signifikan untuk pendapatan. Temuan ini memberikan wawasan baru yang berharga bagi strategi pemasaran di pusat perbelanjaan.

Tujuan: Penelitian ini bertujuan untuk mengevaluasi efektivitas algoritma DBSCAN dalam segmentasi pelanggan mall berdasarkan data demografis dan perilaku belanja mereka. Hasil dari penelitian ini diharapkan dapat memberikan rekomendasi yang lebih baik bagi pengelola mall untuk menentukan strategi pemasaran yang lebih efektif.

Metode/Desain/pendekatan penelitian: Penelitian ini menggunakan algoritma DBSCAN untuk mengelompokkan data pelanggan yang terdiri dari variabel-variabel seperti usia, jenis kelamin, pendapatan tahunan, dan skor pengeluaran. Proses penelitian meliputi tahapan-tahapan seperti pengumpulan data, preprocessing, penentuan parameter DBSCAN, evaluasi klaster, serta visualisasi hasil.

Hasil/Temuan: Hasil penelitian menunjukkan bahwa algoritma DBSCAN mampu membentuk lima cluster utama serta satu cluster outlier. Parameter optimal ditemukan pada nilai epsilon 12,5 dan min_samples 4. Hasil evaluasi menggunakan Silhouette Score juga menunjukkan bahwa kombinasi parameter ini memberikan hasil klasterisasi yang paling baik.

Kebaruan/Kesimpulan: Penelitian ini memberikan bukti bahwa DBSCAN unggul dalam mendeteksi cluster dan outlier pada dataset pelanggan mall. Dengan kemampuan algoritma untuk menangani variasi kepadatan data, hasil penelitian ini dapat membantu pengelola mall dalam menyusun strategi pemasaran yang lebih personal dan sesuai dengan segmen pelanggan.

Kata Kunci : DBSCAN, Segmentasi Pelanggan, Klasterisasi, Data Mall

Introduction

Dengan pesatnya perkembangan teknologi, berbagai sektor seperti korporasi, kesehatan, pendidikan, pemasaran, dan pemerintahan telah mengalami perubahan yang signifikan [20]. Dalam konteks bisnis, terutama pusat perbelanjaan (mall), inovasi merupakan suatu keharusan agar strategi pemasaran dan penjualan dapat tercapai sesuai dengan target [21]. Mall modern tidak hanya berfungsi sebagai tempat transaksi ekonomi, tetapi juga sebagai pusat aktivitas sosial masyarakat, menjadikannya elemen penting dalam kehidupan sehari-hari [4]. Oleh karena itu, sangat penting bagi pengelola mall untuk mengidentifikasi serta menanggapi kebutuhan dan preferensi konsumen dengan cara yang lebih efektif guna mencapai tujuan bisnis yang optimal [15].

[1]*Yusfi Syawali.
Alamat email: yusfisyawali@gmail.com (Syawali), arnita@unimed.ac.id (Arnita), muhammadhaikalhafiz28@gmail.com (Rangkuti), kakamayadi9@gmail.com (Mayadi), silvipratiwi756@gmail.com (Pratiwi), mhdfachrurrazi09@gmail.com (Razi)

Dalam merumuskan kebijakan dan strategi pemasaran, pengelola mall perlu mempertimbangkan sejumlah faktor seperti tingkat pendapatan masyarakat, usia konsumen yang aktif berbelanja, pengeluaran per transaksi, serta aspek demografis lainnya seperti jenis kelamin [15]. Analisis yang komprehensif terhadap variabel-variabel ini memungkinkan pengelola untuk mendapatkan wawasan yang lebih mendalam terkait pola konsumsi dan preferensi pelanggan, yang pada akhirnya dapat digunakan untuk meningkatkan kepuasan konsumen dan mendongkrak penjualan.

Salah satu pendekatan yang dapat diterapkan dalam menganalisis data pelanggan adalah teknik pengelompokan (clustering) dalam data mining. Salah satu algoritma yang sering digunakan adalah DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Algoritma DBSCAN bekerja dengan mengelompokkan data berdasarkan kepadatan objek di sekitarnya dan memiliki keunggulan dalam mendeteksi outlier serta menangani data dengan distribusi yang tidak beraturan tanpa memerlukan penentuan jumlah cluster pada awal analisis [2]. Fleksibilitas DBSCAN dalam mengelola data yang memiliki kepadatan yang bervariasi, serta kemampuannya mendeteksi outlier secara efektif, menjadikannya alat yang sangat berguna dalam analisis data yang kompleks.

Selain DBSCAN, algoritma K-Means merupakan metode clustering yang sederhana dan efisien, terutama untuk data dengan ukuran besar dan distribusi yang relatif seragam [9]. K-Means membagi data ke dalam sejumlah cluster berdasarkan jarak terdekat dari pusat cluster (centroid), sehingga hasil analisisnya dapat diinterpretasikan dengan mudah. Namun, K-Means memiliki keterbatasan dalam mengelola data yang membentuk cluster yang tidak beraturan serta membutuhkan penentuan jumlah cluster sebelum analisis, yang sering kali sulit untuk diperkirakan.

Penelitian yang dilakukan oleh Nur Rohman (2024) membandingkan algoritma K-Means dan K-Medoids dengan menggunakan teknik elbow untuk menentukan jumlah cluster yang optimal. Temuan dari penelitian tersebut menunjukkan bahwa K-Means dengan 5 cluster menghasilkan nilai Silhouette Coefficient tertinggi, yang mengindikasikan kualitas pengelompokan yang lebih baik dibandingkan K-Medoids [8]. Hal ini mempertegas bahwa pemilihan algoritma clustering sangat dipengaruhi oleh karakteristik data yang dianalisis.

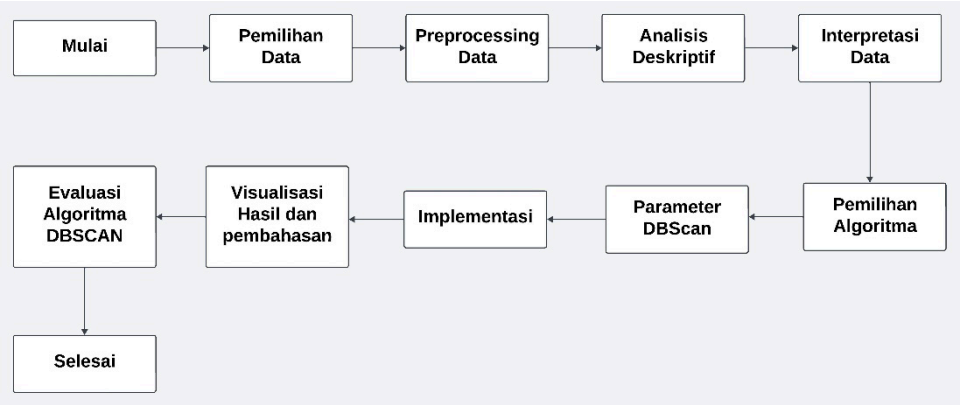
Penelitian ini bertujuan untuk mengevaluasi efektivitas algoritma DBSCAN dibandingkan dengan K-Means dan K-Medoids dalam pengelompokan data pelanggan mall. Melalui perbandingan hasil pengelompokan dari algoritma-algoritma tersebut, penelitian ini bertujuan untuk memberikan rekomendasi yang lebih tepat kepada pengelola mall dalam merumuskan strategi pemasaran yang efektif. DBSCAN diharapkan dapat unggul dalam menangani data dengan distribusi kepadatan yang tidak merata serta mendeteksi outlier, sementara K-Means dan K-Medoids lebih sesuai untuk data dengan distribusi yang lebih seragam dan dapat diukur dengan lebih jelas.

Dengan demikian, hasil penelitian ini diharapkan memberikan kontribusi yang signifikan dalam pengembangan strategi pemasaran berbasis data, serta memberikan wawasan lebih dalam terkait kekuatan dan kelemahan dari masing-masing algoritma clustering dalam pengolahan data pelanggan.

METODE

A. Tahapan Penelitian

Pada penelitian ini memiliki tahapan-tahapan seperti pemilihan data, preprocessing data, pemilihan algoritma, penentuan parameter, implementasi, Evaluasi klastering, Visualisasi hasil, Analisis hasil.



Gambar 1. Tahapan penelitian

B. Pemilihan Data

Penelitian ini menggunakan dataset yang diperoleh dari Kaggel yaitu "Mall Customer Segmentation" memiliki informasi tentang pelanggan dari pusat perbelanjaan. Data terdiri dari 200 entri dengan 5 kolom utama yang mencakup informasi-informasi demografis dan perilaku pelanggan, berikut kolom kolom tersebut adalah:

1. **CustomerID**: Merupakan identifikasi unik bagi setiap pelanggan dalam dataset.
2. **Gender**: Jenis kelamin pelanggan, terdiri dari dua kategori yaitu Male (Laki-laki) dan Female (Perempuan).
3. **Age**: Usia pelanggan, yang diukur dalam tahun.
4. **Annual Income (k$)**: Pendapatan tahunan pelanggan yang dinyatakan dalam ribuan dolar.
5. **Spending Score (1-100)**: Skor belanja pelanggan, diukur berdasarkan tingkat kepuasan pelanggan dengan skala 1 hingga 100.

Berikut di lampirkan statistic deskriptif dari dataset yang digunakan:

	CostumerID	Age	Annual Income	Spending Score (1-100)
Count	200	200	200	200
Mean (Rata-rata)	100.50	38.85	60.56	50.20
Std (Standar Deviasi)	57.88	13.97	26.26	25.82
Min (Nilai Minimum)	1.00	18.00	15.00	1.00
25% (Kuartil Pertama)	50.75	28.75	41.50	34.75
50% (Median/Kuartil Kedua)	100.50	36.00	61.50	50.00
75% (Kuartil Ketiga)	150.25	49.00	78.00	73.00
Max (Nilai Maksimum)	200.00	70.00	137.00	99.00

Tabel 1. Statistic deskriptif

C. Preprocessing Data

Dalam kajian literatur Syaiful menjelaskan bahwa Pembersihan data adalah langkah penting dalam rekayasa data untuk memastikan data bebas dari kesalahan, inkonsistensi, dan ketidakakuratan. Proses ini

melibatkan teknik-teknik seperti deteksi dan penghapusan outlier, imputasi data yang hilang, normalisasi, serta standarisasi data [17]. Pembersihan data yang komprehensif sangat penting untuk menjamin keandalan data sebelum digunakan dalam analisis lebih lanjut. Selain itu pre-processing data mencakup penanganan nilai yang hilang, penghapusan variabel yang tidak diperlukan, visualisasi data, deteksi outlier, serta feature engineering. Feature engineering adalah teknik yang diterapkan setelah data dikumpulkan dan dibersihkan, dan dilakukan sebelum membangun model machine learning [19].

```
mall_data.isnull().sum()
```

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

dtype: int64

Gambar 2. Pengecekan Data Hilang

Penjelasan pada **Gambar 2**, kode `mall_data.isnull().sum()` digunakan untuk memeriksa dan menghitung jumlah nilai yang hilang dalam dataset `mall_data`. Fungsi ini menghasilkan sebuah Series yang menunjukkan jumlah nilai hilang pada setiap kolom. Dalam kasus ini, semua nilai pada Series adalah 0, yang berarti tidak ada nilai yang hilang pada kolom CustomerID, Gender, Age, Annual Income (k$), dan Spending Score (1-100). Dengan demikian, dataset `mall_data` memiliki kualitas data yang baik dan siap untuk analisis lebih lanjut tanpa perlu imputasi atau penghapusan data.

D. Analisis Deskriptif

Setelah di pastikan bahwa tidak ada nilai kosong dalam data, Langkah yang harus di lakukan adalah melakukan analisis deskriptif. Analisis deskriptif digunakan untuk menggambarkan karakteristik dasar dari data yang ada. Ini mencakup perhitungan statistik dasar seperti mean (rata-rata), median (nilai tengah), dan modus (nilai yang paling sering muncul) untuk memahami distribusi dan pola data [5].

Deskripsi rinci dari variabel-variabel numerik yang dianalisis:

1. **CustomerID**
   - **CustomerID** adalah identifikasi unik untuk setiap pelanggan dan bersifat non-inferensial. Kolom ini tidak akan digunakan dalam analisis lebih lanjut namun tetap dimasukkan untuk keperluan referensi atau identifikasi.
2. **Age (Usia)**
   - **Rata-rata (Mean)**: Usia rata-rata pelanggan adalah **38.85 tahun**, yang menunjukkan bahwa kebanyakan pelanggan berada dalam rentang usia dewasa menengah.
   - **Standar Deviasi (Std)**: Sebesar **13.97 tahun**, menunjukkan variasi yang cukup besar dalam usia pelanggan.
   - **Rentang Usia**: Usia termuda adalah **18 tahun** dan tertua adalah **70 tahun**, dengan kuartil pertama di **28.75 tahun**, median di **36.00 tahun**, dan kuartil ketiga di **49.00 tahun**.
3. **Annual Income (Pendapatan Tahunan)**
   - **Rata-rata (Mean)**: Pendapatan tahunan rata-rata pelanggan adalah **60.56 ribu dolar**, yang menunjukkan pelanggan dari berbagai tingkat pendapatan.
   - **Standar Deviasi (Std)**: Sebesar **26.26 ribu dolar**, yang menandakan adanya variasi besar dalam pendapatan pelanggan.
   - **Rentang Pendapatan**: Pendapatan pelanggan berkisar antara **15 ribu dolar** hingga **137 ribu dolar**, dengan kuartil pertama di **41.50 ribu dolar**, median di **61.50 ribu dolar**, dan kuartil ketiga di **78 ribu dolar**. Hal ini menunjukkan adanya distribusi pendapatan yang cukup luas.

4. **Spending Score (Skor Belanja)**
    - **Rata-rata (Mean)**: Skor belanja rata-rata pelanggan adalah **50.20**. Karena rentang skor belanja adalah antara 1 hingga 100, ini menunjukkan bahwa pelanggan cenderung berada di tengah-tengah dalam hal kebiasaan belanja mereka.
    - **Standar Deviasi (Std)**: Sebesar **25.82**, yang menunjukkan bahwa terdapat variasi yang cukup besar dalam kebiasaan belanja pelanggan.
    - **Rentang Skor Belanja**: Skor belanja pelanggan berkisar dari **1** (minimum) hingga **99** (maksimum), dengan kuartil pertama di **34.75**, median di **50.00**, dan kuartil ketiga di **73.00**.

E. Interpretasi

- **Distribusi Usia** menunjukkan bahwa pelanggan berada dalam rentang usia yang luas, namun sebagian besar dari mereka berada di usia produktif (sekitar 36-49 tahun).
- **Distribusi Pendapatan** memperlihatkan adanya pelanggan dengan pendapatan tinggi, namun sebagian besar pelanggan memiliki pendapatan antara 41 ribu hingga 78 ribu dolar per tahun.
- **Distribusi Spending Score** menunjukkan bahwa sebagian besar pelanggan memiliki kebiasaan belanja yang moderat hingga tinggi, dengan skor belanja bervariasi dari 1 hingga 99.

F. Pemilihan Algoritma

Algoritma DBSCAN dipilih karena kemampuannya dalam mengelompokkan data berdasarkan kerapatan (density) dan mengidentifikasi outlier secara otomatis, tanpa perlu menentukan jumlah klaster di awal seperti K-Means. Keunggulan DBSCAN mencakup fleksibilitas dalam menangani bentuk klaster yang tidak teratur dan deteksi noise, sementara kekurangannya adalah ketergantungan pada parameter epsilon dan minimal points, yang dapat mempengaruhi hasil klastering [14].

G. Parameter DBSCAN

DBSCAN adalah metode pengelompokan yang didasarkan pada kepadatan data, yaitu jumlah data (minimum points) yang berada dalam radius Eps (ε) dari setiap titik data [6]. Algoritma ini melakukan pengelompokan berdasarkan dua parameter utama: epsilon dan minpts, yang mempengaruhi jumlah cluster yang dihasilkan [16]. Jarak Euclidean digunakan untuk menghitung jarak antara titik data dengan centroid (C) dalam proses pengelompokan DBSCAN [1].

Proses pengelompokan data menggunakan Algoritma DBSCAN melibatkan beberapa langkah, yaitu:

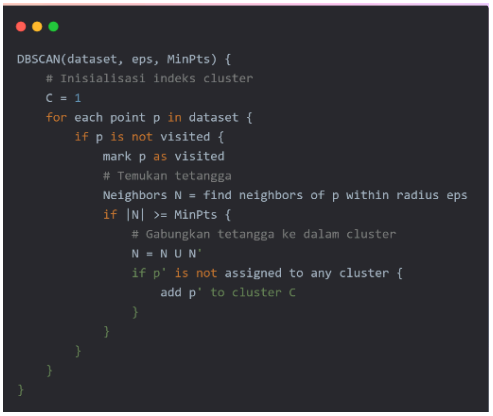
a. Menetapkan nilai Minimal Points (minPts) dan Epsilon (eps).
b. Memilih titik awal secara acak.
c. Menghitung jarak antar titik menggunakan fungsi jarak Euclidean, dengan persamaan berikut:

$$E(x.y) = \sqrt{\sum_{i=0}^{n} \left(x_i - y_i\right)^2} \quad (1)$$

Keterangan :
E (X, Y) = jarak antara titik Xi dengan titik Yi
Xj = nilai titik 1 pada cluster ke-j
Yi = nilai centroid I pada cluster ke-j

d. Cluster terbentuk berdasarkan kepadatan data.
e. Jika jumlah titik dalam radius eps mencapai atau melebihi minpts, maka titik tersebut dikategorikan sebagai titik inti, dan cluster akan terbentuk.
f. Jika titik tersebut adalah titik batas dan tidak ada titik lain dalam jangkauan kepadatan yang dapat dijangkau, proses akan berlanjut ke titik lainnya.
g. Langkah-langkah c hingga f akan diulang sampai semua titik telah diproses.

Pseudocode untuk proses pengelompokan menggunakan algoritma DBSCAN adalah sebagai berikut:



```
DBSCAN(dataset, eps, MinPts) {
    # Inisialisasi indeks cluster
    C = 1
    for each point p in dataset {
        if p is not visited {
            mark p as visited
            # Temukan tetangga
            Neighbors N = find neighbors of p within radius eps
            if |N| >= MinPts {
                # Gabungkan tetangga ke dalam cluster
                N = N U N'
                if p' is not assigned to any cluster {
                    add p' to cluster C
                }
            }
        }
    }
}
```

Gambar 3. Pseudocode DBSCAN

H. Implementasi

Segmentasi pelanggan menggunakan DBSCAN dapat memberikan wawasan tentang kebiasaan dan kebutuhan pelanggan. Sebagai contoh, dalam penelitian yang menerapkan metode RFM (Recency, Frequency, Monetary), DBSCAN dapat digunakan untuk mengelompokkan pelanggan berdasarkan pola perilaku dan nilai transaksi mereka. Dengan kemampuan DBSCAN untuk mengidentifikasi outlier yang menunjukkan perilaku atau nilai transaksi yang tidak biasa, segmentasi pelanggan dapat dilakukan dengan lebih akurat [11].

I. Visualisasi Hasil dan Analisis

Visualisasi Hasil: Pada tahap ini, hasil clustering divisualisasikan untuk memberikan pemahaman visual tentang struktur data yang telah terbentuk. DBSCAN menghasilkan cluster yang divisualisasikan dalam bentuk scatter plot, di mana setiap titik mewakili satu data dan diberi warna berbeda berdasarkan cluster tempat data tersebut berada. Selain itu, outliers yang tidak masuk ke dalam cluster tertentu akan diberi label khusus atau warna berbeda untuk membedakannya dari data yang termasuk dalam cluster. Teknik reduksi dimensi seperti PCA sering digunakan untuk memudahkan visualisasi pada data berdimensi tinggi [7].

Analisis Hasil: Analisis hasil clustering meliputi pengamatan terhadap distribusi cluster dan jumlah cluster yang terbentuk. DBSCAN mampu membentuk cluster berdasarkan densitas data, sehingga ukuran dan jumlah cluster dapat bervariasi tergantung pada distribusi data. Selain itu, DBSCAN mampu mendeteksi outliers yang tidak termasuk ke dalam cluster manapun. Outliers ini dapat memberikan informasi tambahan mengenai titik-titik data yang berbeda secara signifikan dari pola umum [3]. Kualitas cluster dapat dievaluasi menggunakan metode seperti silhouette score, yang membantu dalam mengukur seberapa baik data terkelompokkan dalam cluster [18].

J. Evaluasi Algoritma DBSCAN

Dalam evaluasi algoritma DBSCAN untuk segmentasi pelanggan mall, beberapa metode evaluasi yang dapat dipertimbangkan adalah sebagai berikut:Silhouette Score merupakan metode yang umum digunakan untuk menilai kualitas clustering. Skor ini berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa titik data berada dalam cluster yang lebih kompak dan terpisah dari cluster lain. Dalam penelitian Anda, penggunaan Silhouette Score dapat memberikan wawasan tentang bagaimana nilai ini berubah dengan variasi parameter DBSCAN, seperti nilai epsilon dan MinPts. Sebagai contoh, hasil clustering dengan nilai epsilon 10 dan MinPts 6 menghasilkan Silhouette Score terbaik sebesar 0,521, sedangkan kombinasi lain dapat menunjukkan hasil yang kurang optimal.

Evaluasi dengan Plot Elbow untuk Parameter Tuning juga dapat diterapkan meskipun metode ini lebih sering digunakan dalam konteks K-Means. Dengan menyesuaikan pendekatan ini untuk parameter epsilon dalam DBSCAN, Anda dapat mengidentifikasi titik di mana perubahan jumlah cluster atau outlier mulai melambat. Hal ini akan membantu dalam menentukan nilai epsilon yang optimal untuk analisis lebih lanjut [10].

HASIL DAN PEMBAHASAN

A. Deskripsi Data

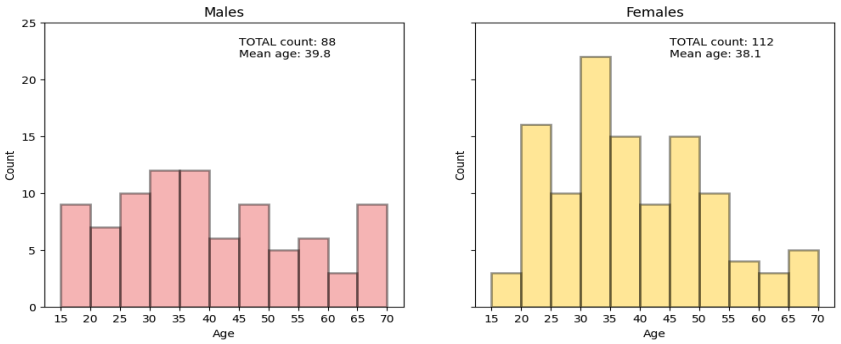
Dataset yang digunakan dalam analisis ini terdiri dari 200 entri yang mencakup lima kolom: *CustomerID*, *Gender*, *Age*, *Annual Income (k$)*, dan *Spending Score (1-100)*. Setelah melakukan pengecekan idak ada nilai yang hilang dalam dataset ini, sehingga semua variabel tersedia untuk analisis lebih lanjut. Variabel *CustomerID* adalah variabel numerik yang merepresentasikan nomor unik untuk setiap pelanggan. Variabel *Gender* adalah variabel kategorikal biner yang menunjukkan jenis kelamin pelanggan (Pria atau Wanita). Sedangkan variabel *Age*, *Annual Income (k$)*, dan *Spending Score (1-100)* adalah variabel numerik yang diukur dalam skala interval.

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

Gambar 4. Describe Data

Secara statistik deskriptif, rata-rata usia pelanggan adalah 38,85 tahun dengan standar deviasi 13,97 tahun, menunjukkan bahwa sebagian besar pelanggan berusia di antara 25 hingga 52 tahun. Pendapatan tahunan rata-rata adalah 60,56 ribu dolar dengan standar deviasi 26,26 ribu dolar, mengindikasikan variasi yang cukup lebar dalam pendapatan tahunan pelanggan. Sementara itu, skor pengeluaran pelanggan rata-rata adalah 50,2 dengan standar deviasi 25,82, yang menunjukkan bahwa pola pengeluaran pelanggan bervariasi dari sangat hemat hingga sangat boros.

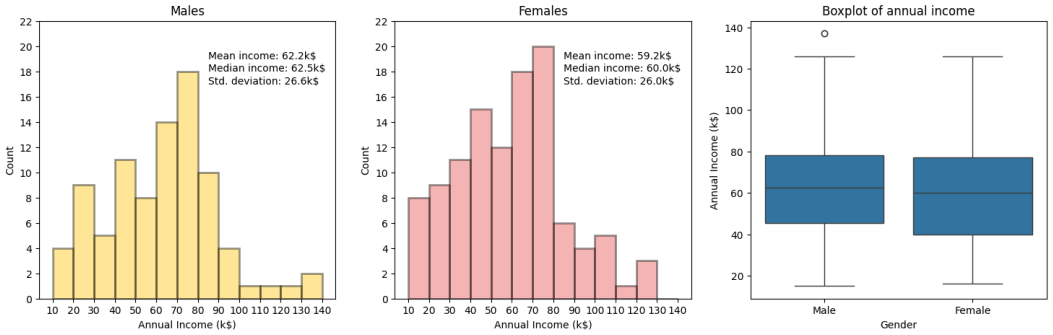
1.1 Distribusi variabel



Gambar 5. Distribusi Usia berdasarkan Jenis Kelamin

Distribusi variabel numerik dianalisis lebih lanjut melalui visualisasi histogram. Hasilnya menunjukkan bahwa rata-rata usia pelanggan pria (39,8 tahun) sedikit lebih tinggi dibandingkan dengan pelanggan wanita (38,1 tahun). Selain itu, distribusi usia pria lebih seragam dibandingkan dengan wanita, di mana kelompok usia terbesar untuk wanita adalah 30-35 tahun. Meskipun ada perbedaan yang teramati dalam distribusi usia antara pria dan wanita, hasil uji Kolmogorov-Smirnov menunjukkan bahwa perbedaan ini tidak signifikan secara statistik dengan kolgomorov-Smirnov test p-value: 0.49.

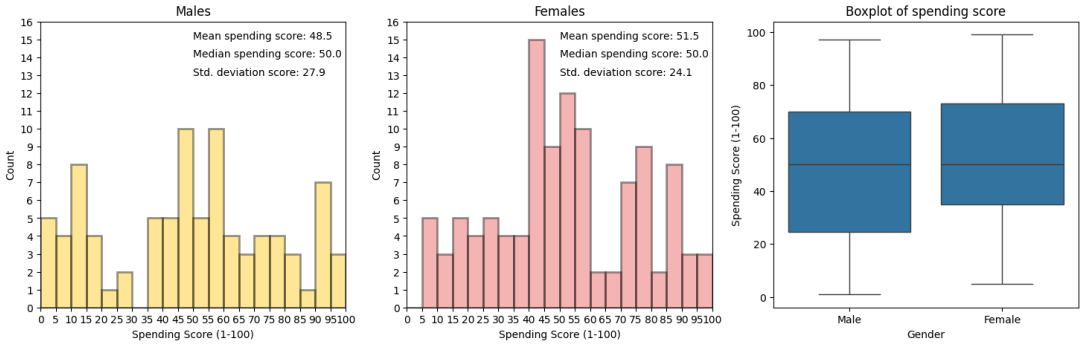
1.2 Analisis Distribusi Data



Gambar 6. Distribusi Pendapatan Tahunan berdasarkan Jenis Kelamin

Grafik distribusi pendapatan tahunan di atas menunjukkan bahwa pelanggan pria memiliki rata-rata pendapatan tahunan sebesar 62,2 ribu dolar, sedikit lebih tinggi dibandingkan dengan pelanggan wanita, yang memiliki rata-rata pendapatan tahunan sebesar 59,2 ribu dolar. Grafik histogram menunjukkan bahwa distribusi pendapatan untuk kedua jenis kelamin serupa, dengan kebanyakan pelanggan memiliki pendapatan antara 40 dan 80 ribu dolar. Namun, kelompok pelanggan pria dengan pendapatan di atas 120 ribu dolar adalah unik.

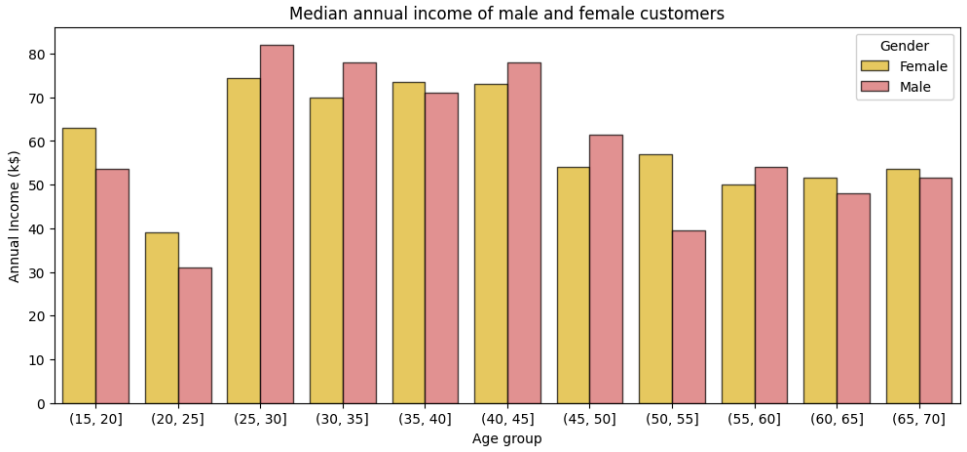
Meskipun data rata-rata pendapatan tahunan antara pria dan wanita sebanding, visualisasi boxplot pendapatan tahunan menunjukkan bahwa pelanggan pria memiliki pendapatan yang sangat tinggi. Ini menunjukkan bahwa, meskipun rata-rata pendapatan tahunan antara pria dan wanita sebanding, ada perbedaan yang signifikan dalam distribusi pendapatan antara pelanggan pria.



Gambar 7. Distribusi Skor Pengeluaran berdasarkan Jenis Kelamin

Selain itu, analisis distribusi skor pengeluaran menunjukkan bahwa pelanggan pria memiliki skor rata-rata 48,5 dan pelanggan wanita memiliki skor rata-rata 51,5. Perbedaan ini menunjukkan bahwa, secara keseluruhan, pelanggan wanita cenderung mengeluarkan lebih banyak uang daripada pelanggan pria. Histogram skor pengeluaran menunjukkan distribusi yang lebih luas pada pelanggan pria, sedangkan pada pelanggan wanita, skor pengeluaran lebih terkonsentrasi di sekitar median (50). Hasil ini didukung oleh boxplot skor pengeluaran, yang menunjukkan bahwa persebaran skor pengeluaran wanita lebih simetris dibandingkan dengan pria.

1.3 Distribusi Pendapatan Berdasarkan Kelompok Usia

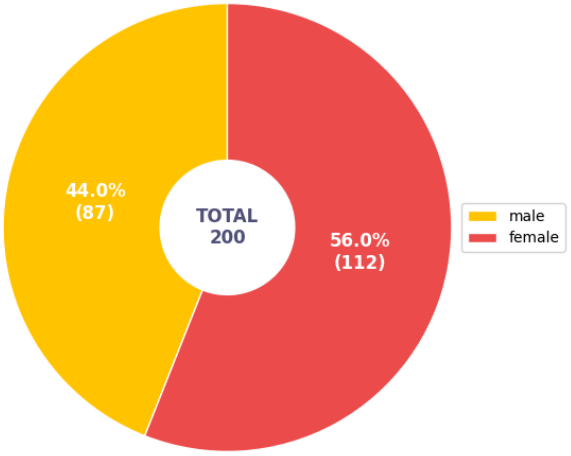


Gambar 8. Median Pendapatan Tahunan berdasarkan Jenis Kelamin dan Kelompok Usia

Sebuah grafik batang yang menunjukkan median pendapatan tahunan pelanggan laki-laki dan perempuan berdasarkan kelompok usia menunjukkan beberapa perbedaan signifikan. Di antara pelanggan wanita, kelompok usia 35-40 tahun memiliki pendapatan median tertinggi, sedangkan di antara pria, kelompok usia 30-35 tahun memiliki pendapatan median tertinggi. Secara umum, hingga usia tertentu (40-45 tahun), pendapatan tahunan meningkat, tetapi kemudian turun setelah itu. Ini menunjukkan bahwa usia tertentu mungkin terkait dengan potensi pendapatan yang lebih tinggi, dan variasi pendapatan tahunan dipengaruhi oleh jenis kelamin dan kelompok usia.

B. Visualisasi Data dan Karakteristik Pelanggan

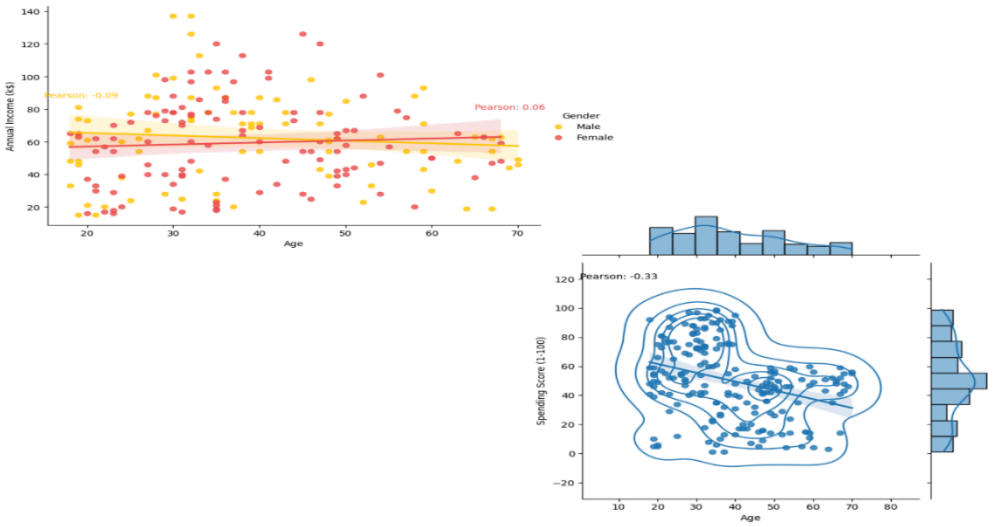
2.1 Analisis Kualitatif



Gambar 9. Jumlah pelanggan perempuan sedikit lebih banyak daripada pelanggan laki-laki

Hasil analisis menggunakan grafik lingkaran/piechart menunjukkan bahwa jumlah pelanggan wanita sedikit lebih besar daripada pria; 112 pelanggan wanita menempati 56% dari total pelanggan. Ditunjukkan oleh perbedaan jumlah ini, ada indikasi awal tentang kemungkinan pembagian pelanggan berdasarkan jenis kelamin.

2.2 Analisis Kuantitatif



Gambar 10. Korelasi antara variabel numerik

Gambar yang terdiri dari dua bagian ini menunjukkan hubungan antara usia pelanggan dengan dua variabel numerik lainnya: pendapatan tahunan (Annual Income) dan skor pengeluaran.

a) Hubungan antara Usia dan Pendapatan Tahunan: Grafik scatter plot di sebelah kiri menunjukkan korelasi antara usia pelanggan dan pendapatan tahunan; titik-titik data dipisahkan menurut jenis kelamin, dengan titik-titik wanita ditunjukkan dengan warna merah dan titik-titik pria ditunjukkan dengan warna kuning. Hasil analisis korelasi Pearson menunjukkan bahwa usia pelanggan pria memiliki korelasi -0,09 dengan pendapatan tahunan wanita 0,06; korelasi ini sangat lemah dan hampir nol, menunjukkan bahwa tidak ada hubungan linier yang signifikan antara usia pelanggan pria dan wanita. Temuan bahwa usia bukanlah prediktor yang baik untuk pendapatan tahunan dalam dataset ini juga diperkuat oleh regresi linier dengan interval keyakinan (ditunjukkan oleh area berwarna).

b) Korelasi antara Usia dan Skor Pengeluaran: Hubungan antara usia dan skor pengeluaran ditunjukkan dalam grafik contour plot di sebelah kanan. Hasil analisis korelasi Pearson adalah -0,33, yang menunjukkan korelasi yang lemah antara skor pengeluaran dan usia. Ini menunjukkan bahwa pelanggan yang lebih tua umumnya memiliki skor pengeluaran yang lebih rendah, tetapi hubungan ini tidak cukup kuat untuk menunjukkan pola yang signifikan. Selain itu, plot densitas menunjukkan bahwa kelompok usia tiga puluh hingga tiga puluh tahun memiliki variasi skor pengeluaran yang lebih besar, sedangkan kelompok pelanggan berusia lebih dari lima puluh tahun cenderung memiliki skor pengeluaran yang lebih terkonsentrasi dan lebih rendah.

Dari kedua grafik ini, dapat disimpulkan bahwa, meskipun ada korelasi negatif kecil antara usia dan skor pengeluaran, variabel usia tidak memiliki korelasi yang signifikan dengan pendapatan tahunan. Ini dapat menunjukkan bahwa usia bukanlah faktor utama yang menentukan pendapatan atau pola pengeluaran pelanggan mall ini. Namun, perbedaan kecil dalam skor pengeluaran berdasarkan usia mungkin menjadi indikasi awal bagi mall untuk mempertimbangkan menggunakan pendekatan baru.

C. Analisis DBSCAN:

3.1 Parameter dan Optimasi

eps (epsilon) dan min_samples adalah dua parameter penting untuk analisis menggunakan algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Untuk menemukan kombinasi parameter yang paling cocok, dilakukan penelitian terhadap beberapa kombinasi nilai eps

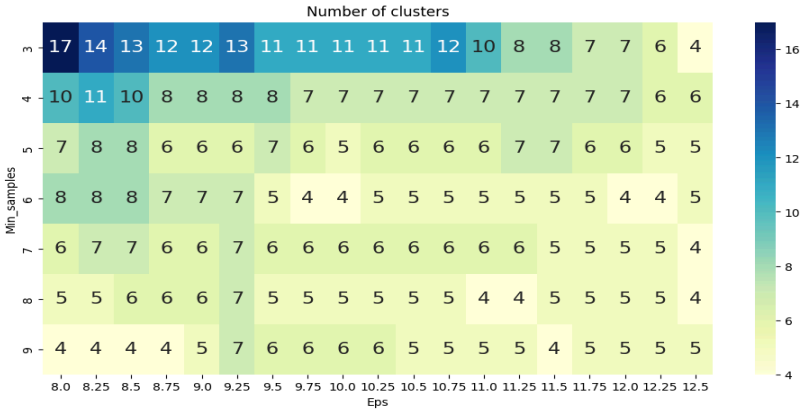
(mulai dari 8 hingga 12.75, dengan kenaikan 0.25) dan min_samples (mulai dari 3 hingga 10). Parameter eps mengatur jarak maksimum antara titik yang dapat dianggap sebagai tetangga.

```
from itertools import product

eps_values = np.arange(8,12.75,0.25) # eps values to be investigated
min_samples = np.arange(3,10) # min_samples values to be investigated

DBSCAN_params = list(product(eps_values, min_samples))
```

Gambar 11. tuning hyperparameter pada algoritma DBSCAN

Hasil eksplorasi ini ditampilkan dalam bentuk heatmap yang menunjukkan jumlah cluster yang terbentuk untuk setiap kombinasi parameter tersebut.

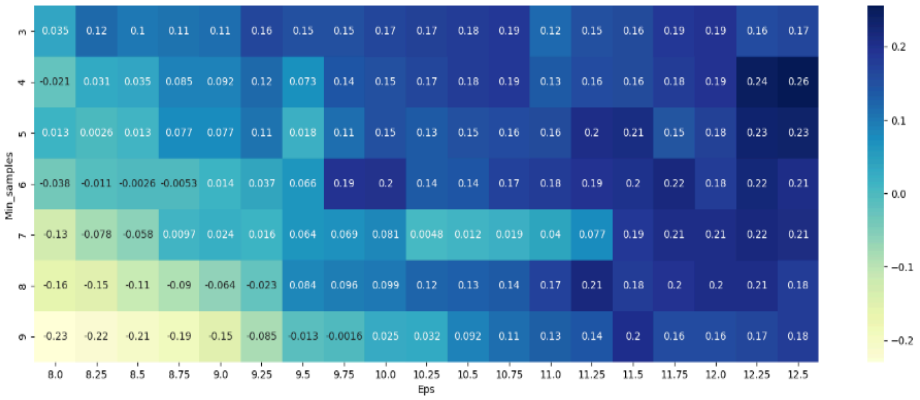
a) Heatmap Jumlah Cluster



Gambar 12. Heatplot

Dilihat dari heatmap di atas, jumlah cluster yang dihasilkan berkisar antara 4 dan 17, tetapi sebagian besar kombinasi parameter menghasilkan cluster antara 4 dan 7. Metrik skor silhouette, yang menunjukkan seberapa baik suatu titik berada di dalam cluster yang ditetapkan, digunakan untuk menentukan kombinasi parameter yang ideal.

b) Heatmap Skor Siluet



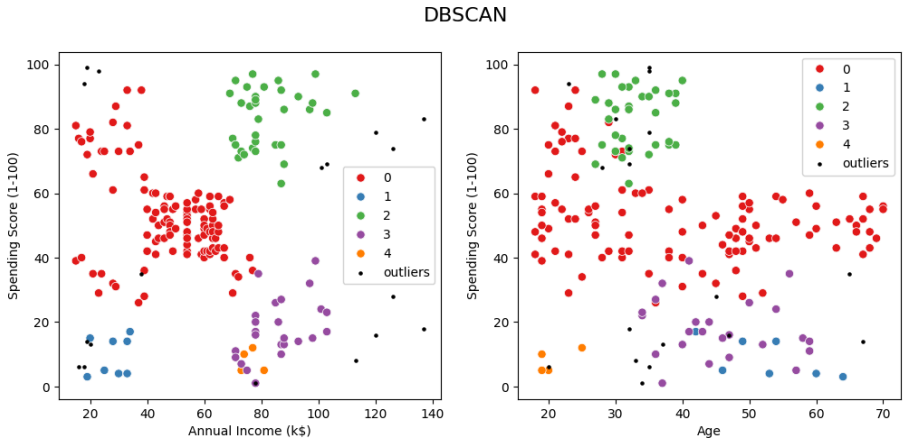
Gambar 13. Heatmap

Untuk setiap kombinasi parameter eps dan min_samples, heatmap kedua menunjukkan skor siluet, yang berkisar antara -1 dan 1, dengan nilai yang lebih tinggi menunjukkan bahwa kluster lebih kompak dan terpisah. Ada eps = 12.5 dan min_samples = 4 untuk kombinasi dengan skor tertinggi (0.26). Akibatnya, kombinasi ini dianggap sebagai parameter yang paling cocok untuk model DBSCAN.

3.2 Plot Hasil Cluster

Model DBSCAN menghasilkan lima cluster utama selain satu cluster yang berisi outlier setelah menentukan parameter terbaik. Jumlah observasi per cluster didistribusikan sebagai berikut:

a) Cluster 0 memiliki 112 observasi
b) Cluster 1 memiliki 8 observasi
c) Cluster 2 memiliki 34 observasi
d) Cluster 3 memiliki 24 observasi
e) Cluster 4 memiliki 4 observasi
f) Outlier (-1) memiliki 18 observasi



Gambar 14. Hasil Clustring

Pendapatan tahunan dan skor pengeluaran, bersama dengan usia dan skor pengeluaran, menunjukkan distribusi pelanggan dalam setiap cluster dalam grafik scatter plot. Semua kelompok ditunjukkan dengan warna yang berbeda, dan outlier ditunjukkan dengan titik hitam. Distribusi cluster ini menunjukkan bahwa beberapa cluster sangat besar (seperti cluster 0), sementara yang lain jauh lebih kecil (seperti cluster 4), yang menunjukkan bahwa ada perbedaan dalam data pelanggan.

3.3 Interpretasi dan Makna

Hasil clustering dengan algoritma DBSCAN adalah grafik scatterplot diataas yang memiliki dua representasi berbeda dari dataset yang sama. Pada grafik kiri, sumbu x menunjukkan pendapatan tahunan dalam ribuan dolar, dan sumbu y menunjukkan skor pengeluaran (1-100). Pada grafik kanan,

sumbu x menunjukkan usia, dan sumbu y menunjukkan skor pengeluaran (1-100). Ini adalah interpretasi dari setiap kelompok:

a) **Cluster 0 (ditandai dengan warna merah):**
   - Pada kedua grafik, cluster ini adalah yang paling besar dan tersebar di berbagai nilai "Annual Income" dan "Age".
   - Cluster ini mungkin mewakili konsumen dengan skor pengeluaran rata-rata hingga tinggi yang memiliki rentang pendapatan atau usia yang lebih luas.

b) **Cluster 1 (ditandai dengan warna biru muda):**
   - Dalam grafik kiri, cluster ini tampaknya mencakup orang-orang dengan pendapatan yang lebih rendah (sekitar 20-40 k$) dan skor pengeluaran rendah.
   - Pada grafik kanan, cluster ini mewakili kelompok usia yang lebih tua dengan skor pengeluaran yang rendah.

c) **Cluster 2 (ditandai dengan warna hijau):**
   - Cluster ini mengelompokkan orang-orang dengan pendapatan yang lebih tinggi (sekitar 60-80 k$) dan skor pengeluaran tinggi pada grafik kiri.
   - Pada grafik kanan, cluster ini mengelompokkan orang-orang dengan skor pengeluaran tinggi tanpa memandang usia.

d) **Cluster 3 (ditandai dengan warna ungu):**
   - Pada grafik kiri, cluster ini mengelompokkan orang-orang dengan pendapatan yang lebih tinggi (sekitar 70-80 k$) tetapi dengan skor pengeluaran rendah.
   - Pada grafik kanan, cluster ini terdiri dari orang-orang berusia di sekitar 40-50 tahun dengan skor pengeluaran rendah.

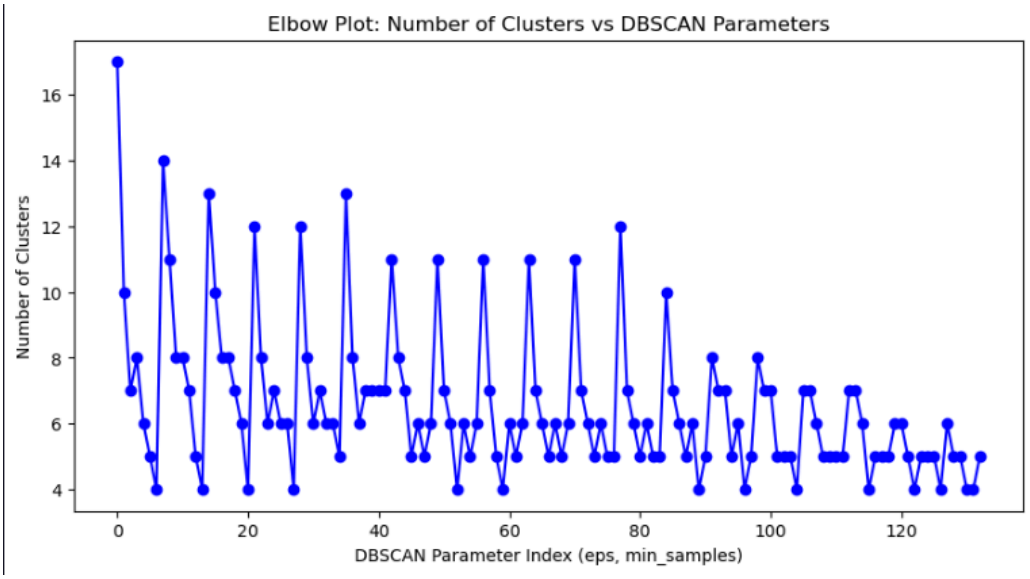
e) **Cluster 4 (ditandai dengan warna oranye):**
   - Dalam kedua grafik, cluster ini tampaknya mengelompokkan individu dengan pendapatan menengah hingga tinggi tetapi dengan skor pengeluaran yang lebih rendah.

f) **Outliers (ditandai dengan warna hitam):**
   - Titik-titik hitam di kedua grafik menunjukkan outlier, yaitu titik data yang tidak sesuai dengan pola cluster yang ada. Outlier ini mungkin mencakup individu dengan karakteristik yang tidak umum atau ekstrem dalam hal pendapatan, usia, atau skor pengeluaran.

Interpretasi ini menunjukkan bagaimana algoritma DBSCAN mengelompokkan data berdasarkan kedekatan titik data dalam ruang multidimensi (pendapatan, usia, dan skor pengeluaran). Cluster yang dihasilkan membantu mengidentifikasi segmen konsumen berdasarkan perilaku pengeluaran mereka, pendapatan tahunan, dan usia.

D. Evaluasi Algoritma DBSCAN:

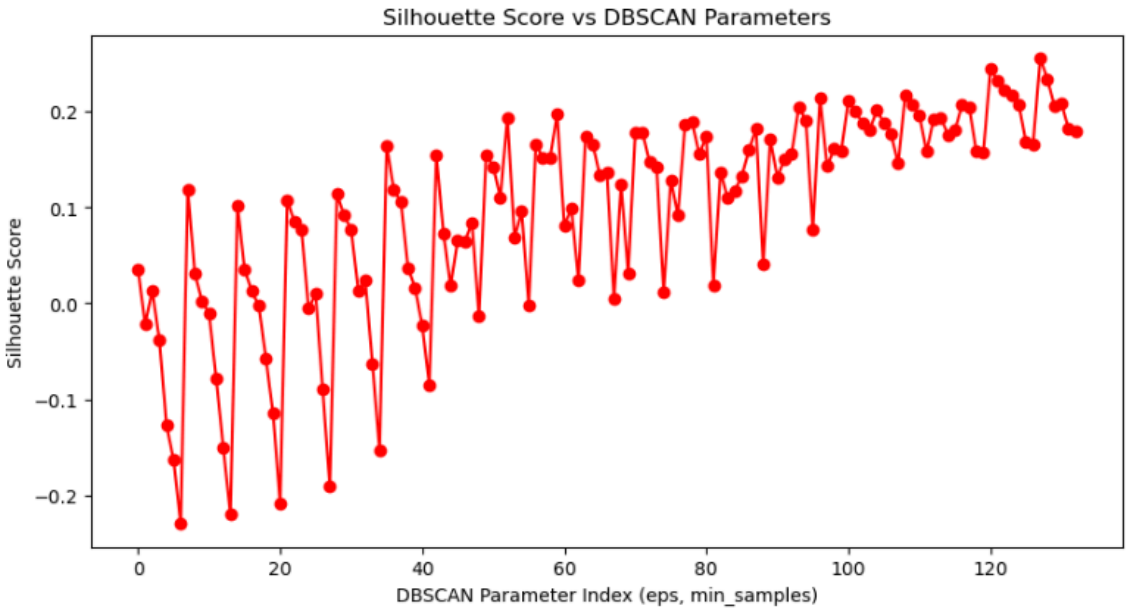
[1] Elbow Plot



Berdasarkan *Elbow Plot* yang Anda berikan, berikut adalah beberapa poin interpretasi mengenai evaluasi DBSCAN:

1. **Jumlah Cluster yang Bervariasi**:
    - Grafik menunjukkan variasi jumlah cluster yang terbentuk berdasarkan perubahan parameter *eps* dan *min_samples*. Nilai *eps* adalah jarak maksimum antar titik yang masih dianggap sebagai tetangga, sementara *min_samples* adalah jumlah minimum titik dalam radius tersebut yang diperlukan untuk membentuk sebuah cluster.
    - Pada indeks parameter rendah (awal), jumlah cluster cenderung tinggi, mencapai hingga 16 cluster.
    - Saat indeks parameter meningkat (kombinasi *eps* dan *min_samples* berubah), jumlah cluster yang terbentuk menurun dan lebih stabil, di sekitar 5 hingga 8 cluster.
2. **Pencarian Titik Elbow**:
    - *Elbow* adalah titik di mana perubahan jumlah cluster mulai melambat dan stabil. Pada grafik ini, ada indikasi bahwa setelah sekitar indeks parameter ke-20 hingga ke-40, jumlah cluster mulai lebih stabil di kisaran 6 hingga 8 cluster.
    - Stabilitas ini bisa menunjukkan kombinasi parameter yang lebih optimal, di mana DBSCAN tidak terlalu agresif membentuk banyak cluster, namun tetap membedakan pola yang signifikan dalam data.
3. **Implikasi**:
    - Titik-titik di mana jumlah cluster mendadak menurun (seperti indeks parameter ke-40 hingga 60) bisa menunjukkan bahwa pada kombinasi *eps* dan *min_samples* tertentu, algoritma DBSCAN lebih sensitif terhadap kepadatan data dan menganggap beberapa titik sebagai *outliers* (label -1), sehingga jumlah cluster yang valid berkurang.
    - Kombinasi parameter di sekitar indeks ke-40 hingga ke-70 tampak menciptakan cluster yang lebih stabil (sekitar 6-8 cluster), yang mungkin memberikan segmentasi yang lebih relevan dan interpretable untuk kasus penggunaan Anda.

[2] Silhouette Score



Grafik tersebut menunjukkan evaluasi algoritma DBSCAN menggunakan *Silhouette Score* terhadap berbagai kombinasi parameter epsilon (eps) dan min_samples, yang diindeks pada sumbu x. Berikut adalah interpretasi dari hasil evaluasi:

1. **Silhouette Score** (sumbu y) adalah ukuran seberapa baik objek dikelompokkan. Nilai positif menunjukkan bahwa objek berada lebih dekat dengan kluster mereka sendiri daripada dengan kluster yang berdekatan, sedangkan nilai negatif menunjukkan bahwa objek mungkin lebih dekat dengan kluster lain. Skor berkisar dari -1 hingga 1.
2. **Kombinasi Parameter DBSCAN**: Sumbu x merepresentasikan indeks parameter kombinasi dari eps dan min_samples. Kombinasi parameter ini sangat mempengaruhi hasil klasterisasi, dan dengan memplot Silhouette Score, kita dapat melihat dampaknya.
3. **Fluktuasi Skor**: Grafik menunjukkan banyak fluktuasi pada skor Silhouette untuk berbagai kombinasi parameter, terutama di awal. Ini menunjukkan bahwa beberapa parameter tidak memberikan hasil klasterisasi yang baik, dengan banyak skor mendekati atau bahkan di bawah 0. Nilai skor yang negatif menandakan klasterisasi yang buruk di mana data mungkin tidak terkelompok dengan baik.
4. **Peningkatan Tren**: Di bagian akhir grafik (indeks lebih dari 60), tampak ada tren peningkatan skor Silhouette yang lebih stabil. Ini menunjukkan bahwa dengan beberapa kombinasi parameter di area tersebut, DBSCAN menghasilkan klaster yang lebih baik, dengan nilai Silhouette mendekati atau di atas 0.2.
5. **Kesimpulan**: DBSCAN bekerja secara efektif hanya dengan kombinasi parameter tertentu. Secara keseluruhan, ada peningkatan yang jelas pada nilai Silhouette untuk beberapa parameter, tetapi ada juga banyak kombinasi parameter yang tidak ideal.

Untuk evaluasi lebih lanjut, Anda bisa memilih kombinasi parameter di mana skor mendekati nilai positif tertinggi, sekitar 0.2 dalam grafik ini.

KESIMPULAN

Berdasarkan dataset yang terdiri dari 200 entri, penelitian ini berhasil menganalisis demografi pelanggan pusat perbelanjaan. Hasil analisis menunjukkan bahwa antara pria dan wanita ada perbedaan yang signifikan dalam pendapatan dan skor pengeluaran. Studi ini menemukan lima cluster utama dan satu cluster outlier dengan menggunakan algoritma DBSCAN. Parameter optimalnya adalah eps = 12,5 dan min_samples = 4. Usia tidak terbukti menjadi indikator yang signifikan untuk pendapatan tahunan, meskipun ada korelasi negatif kecil antara usia dan skor pengeluaran. Hasil ini membantu manajer pusat perbelanjaan membuat strategi pemasaran yang lebih efektif dan menyesuaikan penawaran produk berdasarkan segmen pelanggan yang diidentifikasi.

REFERENSI

[1] David, *Outlier detection in clustering algorithms*. Wiley, 2018.

[2] DQLab, "4 alasan mengapa penting teknik pengolahan data deskriptif untuk analisis data mu," *Belajar Data Science di Rumah*, Dec. 6, 2021. [Online]. Available: https://dqlab.id/4-alasan-mengapa-penting-teknik-pengolahan-data-deskriptif-untuk-analisis-data-mu.

[3] Id, "Modifikasi DBSCAN (Density-Based Spatial Clustering with Noise) Pada Objek 3 Dimensi," *J. Komput. Terapan*, vol. 3, no. 1, pp. 41-52, 2017.

[4] M. Lestari, et al., "Penerapan Tanggap Darurat Pada Pengunjung Salah Satu Mall di Kota Palembang," 2021. [Online]. Available: http://ejournal.uika-bogor.ac.id/index.php/Hearty/issue/archive.

[5] M. T. Anwar, et al., "Wildfire Risk Map Based on DBSCAN Clustering and Cluster Density Evaluation," *Adv. Sustain. Sci. Eng. Technol. (ASSET)*, vol. 1, no. 1, 2019.

[6] N. Ahmed and T. A. Razak, "An overview of various improvements of DBSCAN algorithm in clustering spatial databases," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 2, pp. 360-363, 2016.

[7] N. Rahmah and I. S. Sitanggang, "Determination of optimal epsilon (eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra," in *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 31, no. 1, 2016.

[8] N. Rohman and A. Wibowo, "Perbandingan Metode K-Medoids dan Metode K-Means Dalam Analisis Segmentasi Pelanggan Mall," *SINTECH J.*, vol. 7, no. 1, pp. 49-58, 2024.

[9] N. Sari and A. Primajaya, "Penerapan clustering DBSCAN untuk pertanian padi di kabupaten Karawang," *JIKO (Jurnal Inform. Dan Komputer)* STMIK AKAKOM, vol. 4, no. 1, pp. 28-34, 2019.

[10] P. Y. Utami, S. A. Hudjimartsu, T. A. Viona, and H. Sharfina, "Optimasi parameter algoritma DBSCAN untuk mendeteksi titik panas kebakaran hutan dan lahan," *J. Teknol. Inform.*, vol. 9, no. 3, 2023.

[11] R. John, *Data clustering and analysis*. Academic Press, 2020.

[12] S. Devi, I. K. G. D. Putra, and I. M. Sukarsa, "Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan," *Lontar Komput.: J. Ilm. Teknol. Inf.*, pp. 185-191, 2015.

[13] S. Justine and M. A. Pribadi, "Strategi Komunikasi Pemasaran Pusat Perbelanjaan di Jakarta (Studi Kasus Emporium Pluit Mall)," 2023.

[14] S. Syaiful, R. S. Aminda, and Y. Afrianto, "Influence motor cycle density on noise sound at the highway," *ASTONJADRO*, vol. 12, no. 1, pp. 304-313, 2023.

[15] Saputra and R. Yusuf, "Perbandingan algoritma DBSCAN dan K-MEANS dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM," *MALCOM: Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1346-1361, 2024. [Online]. Available: https://journal.irpi.or.id/index.php/malcom.

[16] Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Trans. Database Syst. (TODS)*, vol. 42, no. 3, pp. 1-21, 2017.

[17] Setiawan, C. Setianingsih, et al., "Clustering Pada Data Sentimen Transportasi Online Menggunakan Algoritma Dbscan," *EProceedings Eng.*, vol. 8, no. 5, 2021.

[18] Smith, *Advanced clustering techniques*. Springer, 2021.

[19] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. V. Broucke, "Special issue on feature engineering editorial," *Mach. Learn.*, 2021. doi: 10.1007/s10994-021-06042-2.

[20] W. Wahyusari and S. Wardani, "Comparison of the K-Means Algorithm and the K-Medoid Algorithm for Clustering UMKM in Kebumen," 2023.

[21] Y. Hapsari, et al., "Analisis Segmentasi Pelanggan Mall Menggunakan Algoritma K-Means," 2023. [Online]. Available: https://www.kaggle.com/code/obrunet/customer-segmentation-k-means.