PERBANDINGAN ALGORITMA *K-MEANS*, *AFFINITY CLUSTERING*, DAN *MINIBATCH K-MEANS* UNTUK ANALISIS SEGMENTASI PASAR

Andrew Castello Purba[1], Teny Handhayani[2]

Program Studi Teknik Informatika, Universitas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail : tenyh@fti.untar.ac.id[2]

Abstrak

Segmentasi pasar adalah proses membagi pasar menjadi kelompok-kelompok pembeli yang homogen berdasarkan karakteristik tertentu. Segmentasi pasar penting bagi perusahaan untuk memahami kebutuhan dan perilaku konsumennya sehingga dapat mengembangkan strategi pemasaran yang lebih efektif. Penelitian ini membandingkan tiga metode *clustering*, yaitu *K-Means Clustering, Affinity Propagation Clustering,* dan *Mini Batch K-Means*, dalam konteks analisis segmentasi pasar. Data yang digunakan adalah dataset *marketing*_campaign.csv yang terdiri dari 29 kolom dan 2240 baris. Eksperimen dilakukan untuk mengevaluasi performa ketiga metode dengan menggunakan metrik siluet. Hasil eksperimen menunjukkan bahwa *Affinity Propagation* menghasilkan nilai *silhouette* tertinggi sebesar 0.5861, diikuti oleh K-Means dengan 0.4675, dan *Mini Batch K-Means* dengan 0.4659. Temuan ini mengindikasikan bahwa *Affinity Propagation* dapat menjadi pilihan yang baik untuk analisis segmentasi pasar pada dataset tersebut. Hasil perbandingan ini memberikan wawasan tentang keunggulan dan kelemahan masing-masing metode *clustering*, membantu pemilih dalam memilih pendekatan yang paling sesuai untuk tujuan analisis mereka. Studi ini memberikan kontribusi pada pemahaman praktis penerapan teknik clustering dalam konteks pemasaran.

Kata kunci : Segmentasi pasar, *K-Means Clustering*, *Affinity Propagation Clustering*, *Mini Batch K-Means*, Nilai siluet

Abstract

Market segmentation is the process of dividing a market into homogeneous groups of buyers based on certain characteristics. Market segmentation is important for businesses to understand the needs and behaviors of their customers so that they can develop more effective marketing strategies.This study compares three clustering methods, namely K-Means Clustering, Affinity Propagation Clustering, and Mini Batch K-Means, in the context of market segmentation analysis. The data used is the marketing_campaign.csv dataset consisting of 29 columns and 2240 rows. Experiments were conducted to evaluate the performance of the three methods using the silhouette metric. The results of the experiments showed that Affinity Propagation produced the highest silhouette value of 0.5861, followed by K-Means with 0.4675, and Mini Batch K-Means with 0.4659. The finding indicates that Affinity Propagation can be a good choice for market segmentation analysis on that dataset. The results of this comparison provide insights into the strengths and weaknesses of each clustering method, helping users choose the approach that is most appropriate for their analytical goals. The study contributes to the practical understanding of the application of clustering techniques in the context of marketing.

Keywords : *Market segmentation, K-Means Clustering, Affinity Propagation Clustering, Mini Batch K-Means, Silhouette score*

1. PENDAHULUAN

Segmentasi pasar adalah proses membagi pasar menjadi kelompok-kelompok pembeli yang homogen berdasarkan karakteristik tertentu [1]. Segmentasi pasar penting bagi perusahaan untuk memahami kebutuhan dan perilaku konsumennya sehingga dapat mengembangkan strategi pemasaran yang lebih efektif.

Dalam upaya untuk memahami dan merespons dinamika pasar yang terus berubah, sebagai peneliti independen, penelitian memfokuskan perhatian pada segmentasi pasar sebagai landasan strategi pemasaran yang efektif. Segmentasi pasar memainkan peran krusial dalam memecah kompleksitas pasar menjadi kelompok pembeli yang homogen berdasarkan karakteristik tertentu, membuka peluang untuk pendekatan pemasaran yang lebih terfokus [2].

Penelitian ini memiliki tujuan untuk membandingkan tiga metode clustering yang dieksplorasi secara pribadi, yaitu : *K-Means Clustering, Affinity Propagation Clustering* [3], dan *Mini Batch K-Means*. Fokus penelitian ini adalah untuk mengevaluasi kinerja relatif ketiganya ketika diterapkan pada dataset yang dikumpulkan, yaitu marketing_campaign.csv, yang mengandung 29 kolom dan 2240 baris informasi yang relevan.

Evaluasi performa metode clustering menjadi fokus utama, dan untuk itu, penelitian ini menggunakan metrik siluet sebagai alat pengukur objektif [4]. Siluet memberikan pemahaman mendalam tentang kemampuan suatu metode membentuk kelompok homogen dan berbeda satu sama lain.

Penelitian ini diharapkan memberikan pemahaman yang lebih mendalam tentang keunggulan dan kelemahan masing-masing metode *clustering* melalui metode *machine learning* yang digunakan. Hasil penelitian ini diharapkan dapat memberikan dasar bagi keputusan yang lebih terinformasi dalam menerapkan teknik *clustering* dalam strategi pemasaran.

Dengan demikian, Research question yang utama dalam penelitian ini adalah:

1. Metode *clustering* apa yang paling optimal untuk analisis segmentasi pasar berdasarkan dataset marketing_campaign.csv?
2. Bagaimana pengaruh variasi jumlah *cluster* pada kinerja algoritma *clustering* dalam segmentasi pasar?
3. Bagaimana pengaruh variasi parameter algoritma *clustering* pada kinerja segmentasi pasar?

2. METODOLOGI

Dalam penelitian ini, metode *clustering* menjadi landasan utama untuk menganalisis segmentasi pasar. Ketiga algoritma utama yang dieksplorasi adalah *K-Means Clustering*, *Affinity Propagation Clustering,* dan *Mini Batch K-Means.* Berikut adalah penjelasan singkat tentang masing-masing metode:

K-Means adalah algoritma *clustering* yang populer dan sederhana. Metode ini membagi data ke dalam k kelompok berdasarkan lokasi pusat massa (*centroid*) setiap kelompok [5]. Iteratif, algoritma ini memperbarui posisi *centroid* hingga konvergensi.

Affinity Propagation memanfaatkan "*propagation*" antara data poin untuk menentukan pusat kelompok. Dalam algoritma ini, setiap poin berkomunikasi dengan yang lain untuk memutuskan pusat kelompok, menangkap struktur data yang kompleks [6].

Mini Batch K-Means adalah varian dari K-Means yang lebih efisien untuk dataset besar [7]. Dibandingkan dengan K-Means tradisional, *Mini Batch K-Means* mengambil sampel acak (*mini-batch*) dari data untuk memperbarui *centroid*, mengurangi waktu komputasi.

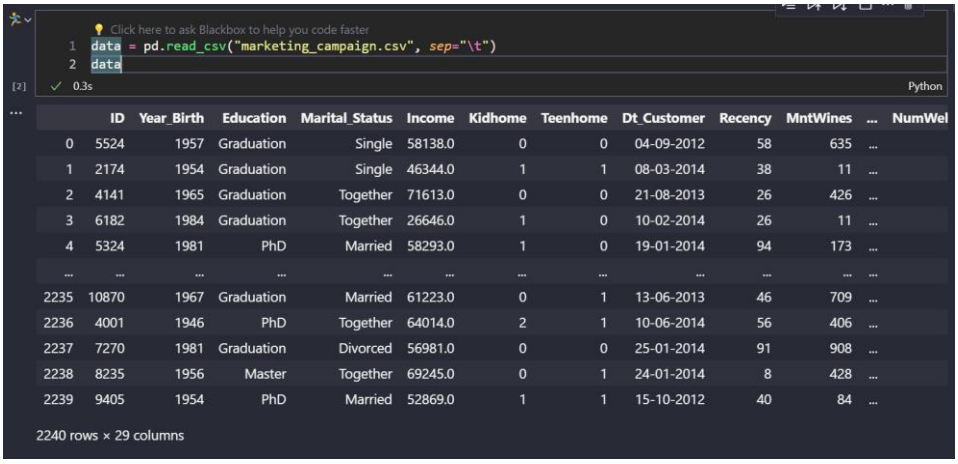
Penerapan metode dilakukan dengan menggunakan dataset marketing_campaign.csv. Evaluasi performa dilakukan dengan metrik siluet, yang mengukur seberapa baik objek dalam satu kelompok dibandingkan dengan kelompok lainnya.

Data yang digunakan dalam penelitian ini diperoleh dari dataset marketing_campaign.csv yang tersedia di Kaggle. Dataset ini mencakup 29 kolom yang mencatat berbagai informasi terkait kampanye pemasaran dan perilaku konsumen, dengan total 2240 baris data, yaitu :

1. ID : Pengidentifikasi unik untuk setiap pelanggan
2. Year_Birth : Tahun lahir pelanggan
3. Education : Tinkat Pendidikan pelanggan
4. Marital_Status : Status pernikahan pelanggan
5. Income : Pendapatan tahunan pelanggan
6. Kidhome : Jumlah anak yang tinggal di rumah bersama pelanggan
7. Teenhome : Jumlah remaja yang tinggal di rumah bersama pelanggan
8. Dt_Customer : Apakah pelanggan menanggapi kampanye pemasaran (variabel target)
9. Recency : Jumlah hari sejak pembelian terakhir pelanggan
10. MntWines : Jumlah uang yang dibelanjakan untuk pembelian anggur dalam dua tahun terakhir
11. MntFruits : Jumlah uang yang dibelanjakan untuk pembelian buah dalam dua tahun terakhir
12. MntMeatProducts : Jumlah uang yang dibelanjakan untuk pembelian produk daging dalam dua tahun terakhir

13. MntFishProducts : Jumlah uang yang dibelanjakan untuk pembelian produk ikan dalam dua tahun terakhir
14. MntSweetProducts : Jumlah uang yang dibelanjakan untuk pembelian produk manis dalam dua tahun terakhir
15. MntGoldProds : Jumlah uang yang dibelanjakan untuk pembelian produk emas dalam dua tahun terakhir
16. NumDealsPurchases : Jumlah pembelian dengan penawaran dalam dua tahun terakhir
17. NumWebPurchases ; Jumlah pembelian melalui web dalam dua tahun terakhir
18. NumCatalogPurchases : Jumlah pembelian katalog dalam dua tahun terakhir
19. NumStorePurchases : Jumlah pembelian toko dalam dua tahun terakhir
20. NumWebVisitsMonth : Rata-rata jumlah kunjungan web per bulan dalam dua tahun terakhir
21. AcceptedCmp3 : Apakah pelanggan menerima kampanye pemasaran 3
22. AcceptedCmp4 : Apakah pelanggan menerima kampanye pemasaran 4
23. AcceptedCmp5 : Apakah pelanggan menerima kampanye pemasaran 5
24. AcceptedCmp1 : Apakah pelanggan menerima kampanye pemasaran 1
25. AcceptedCmp2 : Apakah pelanggan menerima kampanye pemasaran 2
26. Complain : Apakah pelanggan pernah mengajukan keluhan
27. Z_CostContact : Biaya menghubungi pelanggan yang telah distandarisasi
28. Z_Revenue : Pendapatan dari pelanggan yang telah distandarisasi
29. Response : Apakah pelanggan menanggapi kampanye pemasaran (variabel target)

Berikut adalah representasi visual dari data yang diperoleh dari dataset marketing_campaign.csv di Kaggle. Visualisasi ini memberikan berbagai informasi yang terkait dengan kampanye pemasaran yang telah dilaksanakan. Gambar 1 dibawah ini menggambarkan data mentah yang akan digunakan dalam analisis



```
Click here to ask Blackbox to help you code faster
1 data = pd.read_csv("marketing_campaign.csv", sep="\t")
2 data
```

[2] ✓ 0.3s Python

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWel
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	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	10-02-2014	26	11	...	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	
...	...	...	...	...	...	...	...	...	...	...	...	
2235	10870	1967	Graduation	Married	61223.0	0	1	13-06-2013	46	709	...	
2236	4001	1946	PhD	Together	64014.0	2	1	10-06-2014	56	406	...	
2237	7270	1981	Graduation	Divorced	56981.0	0	0	25-01-2014	91	908	...	
2238	8235	1956	Master	Together	69245.0	0	1	24-01-2014	8	428	...	
2239	9405	1954	PhD	Married	52869.0	1	1	15-10-2012	40	84	...	

2240 rows × 29 columns

Gambar 1 Visualisasi Dataset

Data yang dikumpulkan diolah secara awal untuk menghilangkan data yang tidak lengkap atau tidak relevan. Data yang tersisa kemudian disimpan dalam format csv. Kolom-kolom dalam dataset marketing_campaign.csv dapat dikelompokkan menjadi dua kategori, yaitu:

1. Data demografis: kolom-kolom yang mencatat informasi tentang pelanggan, seperti usia, pendidikan, status pernikahan, pendapatan, dan jumlah anak.
2. Data perilaku: kolom-kolom yang mencatat informasi tentang perilaku pembelian pelanggan, seperti jumlah pembelian, jumlah hari sejak pembelian terakhir, dan jenis saluran pembelian.

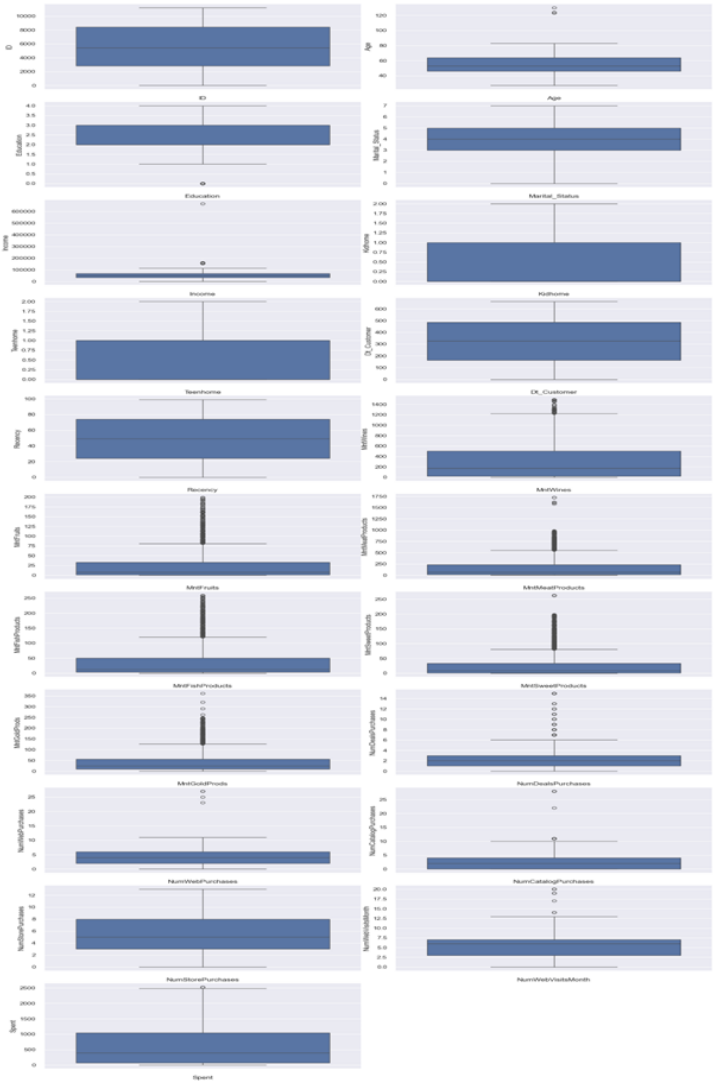
Pengolahan data melibatkan pembersihan data, penanganan nilai-nilai yang hilang, dan penskalaan variabel jika diperlukan [8]. Analisis data mencakup penerapan metode clustering pada dataset dan

pengukuran kinerja metode menggunakan metrik siluet. Proses ini dimulai dengan identifikasi nilai-nilai yang hilang pada dataset marketing_campaign.csv. Setelah identifikasi, langkah-langkah berikut diambil:

Dalam analisis, fokus pada kolom Income yang memiliki nilai hilang pada 2216 dari 2240 baris data. Identifikasi lokasi nilai-nilai yang hilang dilakukan untuk mengevaluasi dampaknya terhadap integritas data dan relevansi analisis segmentasi pasar [9], khususnya terkait pemahaman pendapatan konsumen. Analisis dilakukan untuk mengidentifikasi lokasi dan seberapa signifikan nilai-nilai yang hilang.

Kolom *Education*, Marital Status, Income, Kidhome, Teenhome, NumWebPurchases, NumStorePurchases, *Respone* menunjukkan distribusi yang cenderung normal, dengan sedikit kecondongan. Sementara itu kolom *Recency* memiliki distribusi miring ke kanan, yang menunjukkan kecenderungan nilai yang lebih tinggi dalam sampel data, kemudian kolom NumDealsPurchases, NumCatalogPurchases, NumWebPurchases yang menunjukan kecenderungan nilai yang lebih rendah dalam sampel data.

Pengecekan outlier dillakukan dengan cara memvisualisasikan melalui boxplot dapat memperlihatkan bahwa nilai-nilai ekstrim dalam kolom-kolom yang bersifat kontinu [10]. Namun, perlu dicatat bahwa dalam dataset ini, beberapa kolom memiliki tipe data kategorikal yang tidak cocok untuk visualisasi dengan metode tersebut. Outlier-outlier ini akan ditangani sesuai dengan metode yang sesuai untuk memastikan integritas analisis data [11]. Tindakan ini termasuk dalam upaya untuk mempersiapkan dataset yang tepat untuk penggunaan dalam model *clustering* dan meminimalkan dampak outlier pada hasil analisis [12]. Gambar 2 dibawah ini menampilan visualisasi setiap kolom dengan tipe Numerikal dengan boxplot.



Gambar 2 Hasil Visualisasi Boxplot dari Kolom Numerikal

Setelah melakukan pendeteksian nilai yang hilang, akan dilanjutkan proses untuk menangani nilai-nilai yang hilang pada kolom *Income*, dengan metode mean dikarenakan sebagian besar data yang sudah terisi pada kolom tersebut cenderung miring ke kanan [13]. Menggunakan nilai rata-rata dianggap dapat memberikan representasi yang akurat tanpa menyimpang secara signifikan dari distribusi sebenarnya [14].

Dalam langkah pra-pemrosesan data, normalisasi data dilakukan untuk memastikan bahwa atribut-atribut dalam dataset berada dalam rentang yang serupa [15]. Hal ini bertujuan untuk meningkatkan kinerja algoritma *clustering* dengan meminimalkan dampak variabilitas skala pada hasil prediksi. Ada dua jenis distribusi data yang perlu dinormalisasi, yaitu :

1. Distribusi Normal

   Distribusi normal adalah distribusi data yang berbentuk lonceng. Dalam dataset ini, terdapat tiga atribut yang memiliki distribusi normal, yaitu Dt_Customer, *Recency*, dan *Income*. Untuk atribut-atribut ini, digunakan metode Standard *Scaler* untuk menormalisasinya [16]. Standard *Scaler* adalah metode normalisasi yang paling umum digunakan. Metode ini bekerja dengan cara menghitung rata-rata dan standar deviasi dari data, kemudian membagi setiap nilai data dengan standar deviasi tersebut.
2. Distribusi Miring

   Distribusi miring adalah distribusi data yang tidak berbentuk lonceng [17]. Dalam dataset ini, terdapat lima belas atribut yang memiliki distribusi miring kanan. Tahap ini mengjalankan metode Box-Cox Transformation dan *scaling* Standard Scaler.
3. Normalisasi Variabel Dummy untuk Kolom Kategorikal

   Dalam dataset ini, terdapat lima atribut kategorikal, yaitu Education, Marital_Status, Kidhome, Teenhome, dan Total_Promos. Untuk atribut-atribut tersebut digunakan pendekatan variable dummy atau one-hot encoding [18].

2.4. Metode Evaluasi

Sejumlah metrik evaluasi kinerja model clustering untuk mengevaluasi tiga algoritma yang berbeda, yaitu *K-Means Clustering*, *Affinity Clustering*, dan *Mini Batch K-Means* [19]. Metrik-metrik ini dapat membantu memahami sejauh mana setiap algoritma mampu melakukan clustering dan metode yang terbaik dalam menganalisa segmentasi pasar. Dalam penelitian ini, kami menggunakan tiga metrik evaluasi untuk menilai kinerja algoritma clustering yang diterapkan:

1. Silhouette score mengukur kedekatan antar-sampel dalam satu cluster dan jarak antar-sampel antara dua cluster yang berbeda [20]. Nilai Silhouette score yang tinggi menunjukkan bahwa sampel dalam satu cluster terkelompok dengan baik dan terpisah dari sampel dalam cluster lainnya.
2. Calinski-Harabasz score mengukur perbandingan antara keragaman antar-cluster dan keragaman intra-cluster [21]. Nilai Calinski-Harabasz score yang tinggi menunjukkan bahwa cluster-cluster yang terbentuk memiliki keragaman antar-cluster yang tinggi dan keragaman intra-cluster yang rendah.
3. Davies-Bouldin score mengukur kesenjangan antar-cluster. Nilai Davies-Bouldin score yang rendah menunjukkan bahwa cluster-cluster yang terbentuk tidak saling tumpang tindih [22].

Pemisahan data ini merupakan langkah krusial dalam perbandingan kinerja algoritma, dan hasilnya akan menjadi dasar bagi tahap-tahap berikutnya dalam penelitian [23]. Selanjutnya, tiga algoritma clustering, yaitu *K-Means Clustering*, *Affinity Clustering*, dan *Mini Batch K-Means*, disiapkan untuk pengujian:

1. *K-Means Clustering*: Menggunakan fungsi *KElbowVisualizer* untuk menentukan jumlah *cluster optimal*. Visualisasi ini membantu mengidentifikasi *elbow point* pada kurva inertia, dan hasilnya menunjukkan bahwa cluster optimal adalah 5.
2. *Affinity Clustering*: Tidak menggunakan konsep cluster, melainkan menentukan fungsi damping. Nilai damping faktor yang optimal untuk algoritma *affinity propagation* berada dalam rentang 0.5 hingga 0.9. Nilai damping factor di atas 0.9 dapat menyebabkan konvergensi algoritma yang terlalu cepat, menghasilkan hasil clustering yang suboptimal.
3. *Mini Batch K-Means*: Sama seperti *K-Means Clustering*, menggunakan fungsi KElbowVisualizer untuk menentukan jumlah cluster optimal. Hasilnya menunjukkan bahwa cluster terbaik adalah 5.

Dengan tahapan ini, dilakukan identifikasi parameter kritis, seperti jumlah cluster atau damping factor, yang memainkan peran penting dalam hasil clustering [24]. Tahap berikutnya melibatkan implementasi masing-masing algoritma dengan parameter optimal yang telah ditentukan untuk evaluasi lebih lanjut.

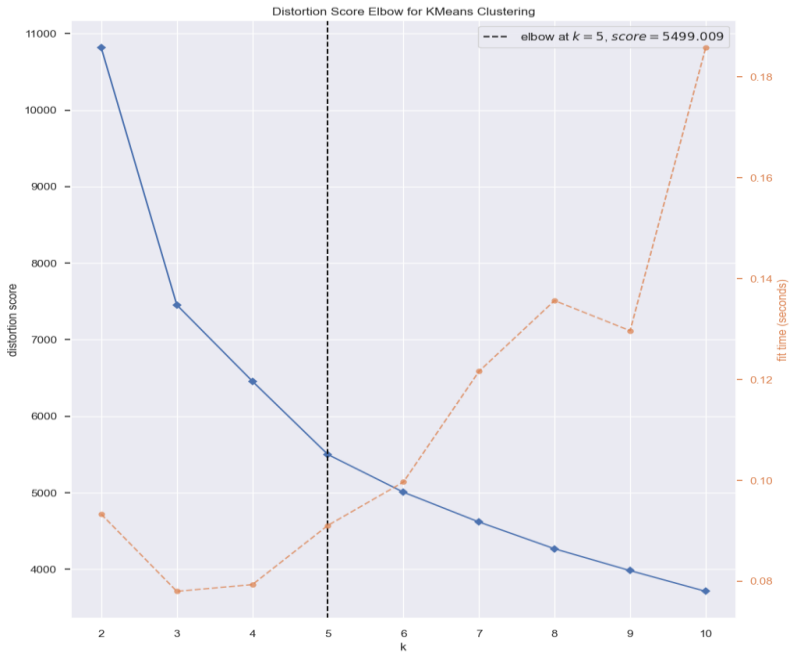
3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan metode *K-Means Clustering*, *Affinity Clustering*, dan *Mini Batch K-Means*, dalam tugas menentukan metode *cluster* yang terbaik dalam segmentasi pasar. Hasil dari masing-masing algoritma ditampilkan pada Tabel 1, 2, dan 3.

Tabel 1 Clustering menggunakan K-Means

Metriks Evaluasi	Skor Evaluasi
Silhouette Score	0.4675
Calinski-Harabasz index	2513.68
Davies-Bouldin index	0.865

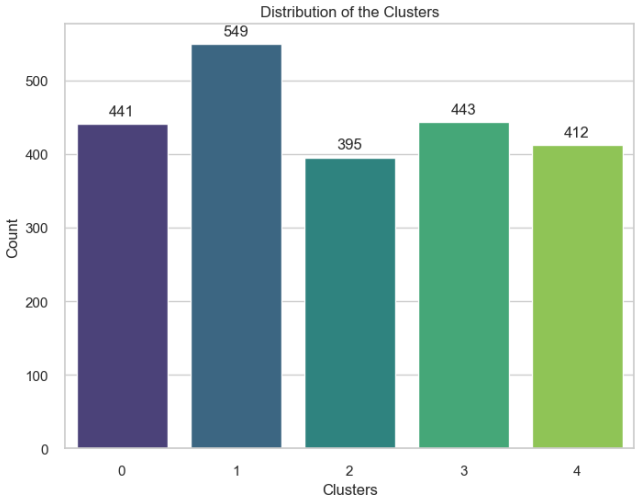
Dalam konteks tugas segmentasi pasar, hasil *clustering* menggunakan *K-Means Clustering* menunjukkan performa yang cukup baik. *Silhouette Score* sebesar 0.4675 mencerminkan tingkat kesamaan yang relatif baik antar-objek dalam kluster. Calinski-Harabasz Index yang tinggi, mencapai 2531.68, mengindikasikan bahwa cluster-cluster yang terbentuk memiliki keragaman antar-cluster yang tinggi dan keragaman *intra-cluster* yang rendah. Hasil ini mendukung efektivitas algoritma *K-Means* dalam memisahkan kelompok pelanggan dengan distribusi yang merata. Davies-Bouldin Index sebesar 0.865 menunjukkan tingkat tumpang tindih yang wajar antar-kluster, menandakan bahwa kluster-keluster yang terbentuk cenderung terpisah satu sama lain tanpa adanya tumpang tindih yang signifikan. Keseluruhan, metrik evaluasi ini memberikan gambaran positif tentang kemampuan *K-Means Clustering* dalam membentuk kelompok pelanggan yang relevan untuk analisis segmentasi pasar. Gambar 3 memperlihatkan performa *K-Means Clustering* dengan berbagai nilai k dalam tugas *clustering* segmentasi pasar.



Gambar 3 KElbow Visualizer K-Means Clustering

Pada grafik tersebut, nilai DBE mulai menurun secara signifikan setelah nilai k mencapai 5. Hal ini menunjukkan bahwa pada nilai k = 5, cluster-cluster yang terbentuk sudah cukup compact dan tidak ada lagi peningkatan signifikan dalam *compactness* jika k dinaikkan lagi. Oleh karena itu, dapat disimpulkan bahwa jumlah *cluster* yang optimal untuk tugas clustering segmentasi pasar tersebut adalah 5. Dengan jumlah cluster 5, data-data dapat dikelompokkan menjadi 5 kelompok yang berbeda berdasarkan kesamaan

karakteristiknya. Gambar 4 memperlihatkan distribusi data setiap cluster menggunakan metode *K-Means Clustering* dalam tugas clustering segmentasi pasar.



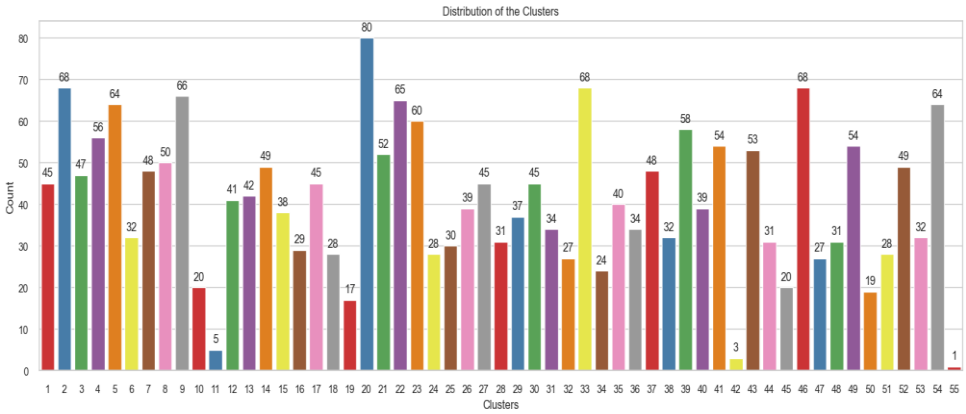
Gambar 4 Distribusi Data *Cluster K-Means Clustering*

Berdasarkan distribusi *cluster K-Means Clustering*, terdapat dua cluster utama yang kemungkinan mewakili segmentasi pasar yang luas. Hal ini menunjukkan bahwa terdapat dua segmentasi pasar yang luas dengan proporsi pelanggan yang signifikan. Selain itu, terdapat tiga cluster yang lebih kecil yang kemungkinan mewakili segmentasi pasar yang lebih spesifik. Keberadaan cluster kecil ini dapat menjadi peluang bagi perusahaan untuk menargetkan strategi marketing dan penjualan yang lebih personal dan terfokus.

Tabel 2 *Clustering* menggunakan *Affinity Clustering*

Metriks Evaluasi	Skor Evaluasi
Silhouette Score	0.5861
Calinski-Harabasz index	9262.96
Davies-Bouldin index	0.6252

Hasil *clustering Affinity Clustering* menunjukkan kinerja yang sangat baik berdasarkan metrik evaluasi yang digunakan. *Silhouette Score* yang tinggi mencerminkan tingkat kesamaan yang optimal antar-objek dalam kluster, sementara *Calinski-Harabasz Index* yang tinggi menunjukkan bahwa cluster-cluster yang terbentuk memiliki keragaman yang sangat baik. *Davies-Bouldin Index* yang rendah menandakan bahwa kluster-keluster cenderung terpisah satu sama lain tanpa adanya tumpang tindih yang signifikan. Dengan demikian, hasil ini memberikan indikasi positif terkait kemampuan *Affinity Clustering* dalam memahami pola dan struktur dalam data pelanggan, menjadi landasan yang kuat untuk analisis lebih lanjut dalam tugas segmentasi clustering. Gambar 5 memperlihatkan distribusi data setiap cluster menggunakan metode *Affinity Clustering* dalam tugas clustering segmentasi pasar.



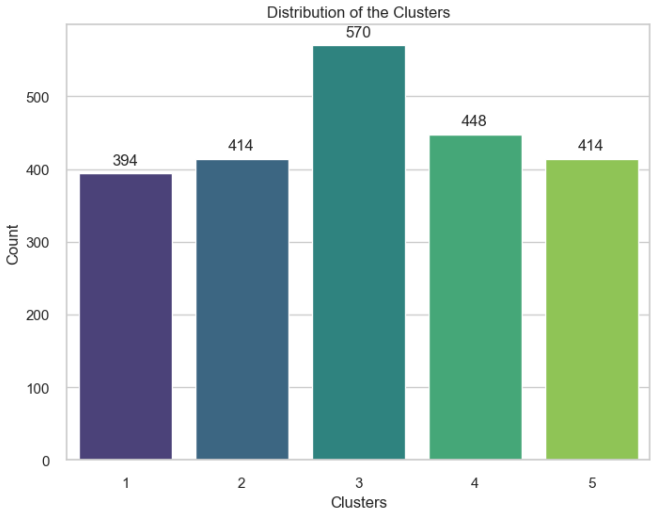
Gambar 5 Distribusi Data *Cluster Affinity Clustering*

Gambar distribusi *cluster Affinity Clustering* menunjukkan beberapa hal menarik. Pertama, terdapat banyak sekali *cluster* berdasarkan hasil visualisasi dibandingkan dengan 2 metode lain. Hal ini menunjukkan kemungkinan adanya pembagian adanya pembagian dan pemisahan kluster dengan sangat baik. Dikarenakan memiliki kompleksitas data yang tinggi, maka diperlukan lebih banyak cluster untuk memodelkan variasi yang ada secara akurat.

Tabel 3 *Clustering* menggunakan *Mini Batch K-Means*

Metriks Evaluasi	Skor Evaluasi
Silhouette Score	0.4659
Calinski-Harabasz index	2536.33
Davies-Bouldin index	0.8763

Hasil clustering menggunakan *Mini Batch K-Means Clustering* menunjukkan kinerja yang relatif baik berdasarkan metrik evaluasi yang digunakan. *Silhouette Score* yang cukup tinggi menandakan bahwa kelompok pelanggan yang terbentuk saling mirip dan terpisah dengan baik, mencerminkan tingkat kesamaan optimal antar-objek dalam kluster. *Calinski-Harabasz Index* yang tinggi menunjukkan bahwa cluster-cluster memiliki keragaman antar-cluster yang tinggi dan keragaman intra-cluster yang rendah, mendukung efektivitas *Mini Batch K-Means* dalam memisahkan kelompok pelanggan dengan distribusi yang merata. *Davies-Bouldin Index* yang rendah menegaskan bahwa kluster-kluster cenderung terpisah satu sama lain tanpa adanya tumpang tindih yang signifikan. Gambar 6 memperlihatkan distribusi data setiap cluster menggunakan metode *Mini Batch K-Means Clustering* dalam tugas *clustering* segmentasi pasar.



Gambar 6 Distribusi Data *Cluster Mini Batch K-Means*

Gambar distribusi *cluster Mini-Batch K-Means Clustering* menunjukkan adanya lima kelompok utama dengan kemungkinan segmentasi yang lebih kecil. Data diatas menunjukkan bahwa kelompok 3 memiliki distribusi data tertinggi diantara yang lain. Dengan hasil ini, *Affinity Clustering* menjadi pilihan yang solid untuk analisis segmentasi *clustering*, memberikan landasan yang kuat untuk langkah-langkah selanjutnya dalam pemahaman perilaku dan preferensi pelanggan.

Dalam mengevaluasi hasil eksperimen *clustering metode K-Means Clustering, Affinity Clustering, dan Mini Batch K-Means*, perbandingan kinerja ketiga algoritma diperoleh melalui metrik evaluasi *Silhouette Score, Calinski-Harabasz Index*, dan *Davies-Bouldin Index*. Berikut adalah rangkuman lengkap hasil dan pembahasan:

1. *K-Means Clustering*:

   Pembahasan: *K-Means Clustering* menghasilkan kinerja yang baik dengan *Silhouette Score* yang positif, menunjukkan tingkat kesamaan yang memadai antar-objek dalam kluster. *Calinski-Harabasz Index* yang tinggi menandakan bahwa cluster-cluster yang terbentuk memiliki keragaman yang baik. Meskipun demikian, perlu diperhatikan bahwa *Davies-Bouldin Index* menunjukkan sedikit tumpang tindih antar-kluster.
2. *Affinity Clustering*:

Pembahasan: *Affinity Clustering* memberikan kinerja yang sangat baik dengan *Silhouette Score* yang paling tinggi dan *Calinski-Harabasz Index* yang mencapai puncak. *Davies-Bouldin Index* yang rendah menegaskan kemampuan algoritma ini dalam membentuk kluster yang terpisah dengan baik.

3. *Mini Batch K-Means*:

   Pembahasan: *Mini Batch K-Means* menunjukkan hasil yang sebanding dengan *K-Means Clustering*, menunjukkan keefektifan dalam membentuk kluster dengan tingkat kesamaan yang optimal dan distribusi yang merata.

Analisis Komparatif:

a. Ketiga algoritma clustering memberikan kinerja yang baik, mampu membentuk kluster dengan tingkat kesamaan yang tinggi dan distribusi yang merata.
b. *Affinity Clustering* menunjukkan hasil yang paling menonjol dikarenakan memberikan hasil metrik evaluasi yang sangat baik, menandakan potensi optimal dalam analisis segmentasi pasar.
c. Kesamaan hasil metrik evaluasi antara *K-Means Clustering* dan *Mini Batch K-Means* menunjukkan konsistensi dan *robustness* dalam pemilihan metode *clustering*.

4. PENUTUP

Dalam penelitian ini, penulis mengkaji tiga metode *clustering*, yaitu *K-Means Clustering*, *Affinity Clustering*, dan *Mini Batch K-Means*, dalam konteks analisis segmentasi pasar. Berdasarkan evaluasi kinerja menggunakan metrik-metrik seperti *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*, dapat disimpulkan bahwa:

1. Kinerja Metode *Clustering*:
   a. *K-Means Clustering* menunjukkan kinerja yang cukup baik dengan kemampuan membentuk kelompok pelanggan dengan tingkat kesamaan yang memadai.
   b. *Affinity Clustering* menonjol dengan hasil evaluasi yang sangat baik, menandakan potensi optimal dalam analisis segmentasi pasar.
2. Keunggulan *Affinity Clustering*:
   a. *Affinity Clustering* memberikan hasil evaluasi yang superior, terutama dengan *Silhouette Score* dan *Calinski-Harabasz Index* yang lebih tinggi.
   b. *Davies-Bouldin Index* yang rendah pada *Affinity Clustering* menunjukkan kemampuan algoritma ini dalam membentuk kluster yang terpisah dengan baik.
3. Konsistensi *Mini Batch K-Means*:
   a. *K-Means Clustering* menunjukkan hasil yang sebanding dengan *Mini Batch K-Means Clustering*, menegaskan keefektifan dalam membentuk kluster dengan tingkat kesamaan yang optimal dan distribusi yang merata.
   b. Kesamaan hasil metrik evaluasi antara *Mini Batch K-Means* dan *Mini Batch K-Means Clustering* menunjukkan konsistensi dan robustness dalam pemilihan metode *clustering*.
4. Rekomendasi Penggunaan Metode:
   a. *Affinity Clustering* dapat dianggap sebagai pilihan yang unggul dalam konteks analisis segmentasi pasar dikarenakan dapat memberikan hasil yang paling memuaskan.
   b. Pemilihan antara keduanya sebaiknya didasarkan pada karakteristik data, kebutuhan analisis, dan tujuan spesifik penelitian.

Dengan demikian, hasil evaluasi kinerja ketiga metode *clustering* memberikan pandangan yang mendalam tentang potensi dan kecocokan masing-masing algoritma dalam mendukung analisis segmentasi pasar. Rekomendasi penggunaan metode tertentu dapat membantu peneliti atau praktisi untuk memilih pendekatan yang paling sesuai dengan konteks dan tujuan analisis mereka.

DAFTAR PUSTAKA

[1] C. Yuan and H. Yang, “Research on k-value selection method of K-Means clustering algorithm,” *Third Internatinal Symposium on Inteligent Information Technology and Security Informatics*, vol. 2, pp. 226-235, 2019.

[2] V. V, "Customer Segmentation Using Machine Learning," *International Conference on Computational Techniques, Electronics and Mechanical Systems(CTEMS),* vol. 08, 2021.

[3] K. Dhiraj, "Implementing Customer Segmentation Using Machine Learning [Beginner Guide]," Hohhot, 2021.

[4] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, 2020.

[5] V.Vijilesh, "CUSTOMER SEGMENTATION USING MACHINE LEARNING," *International Research Journal of Engineering and,* vol. 08, 2021.

[6] Z. Xu, Y. Lu and J. Yu, "Research on Mini-Batch Affinity Propagation Clustering Algorithm," in *IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, Shenzhen, 2022.

[7] M. M. Amini and H. Amini, "A comparison of K-means, Mini-batch K-means, and Gaussian mixture models for customer segmentation," *Journal of Business & Economic Research,* Vols. 17(2), 1-13, 2019.

[8] d. V. V. Werner, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl. Inf. Syst.,* 2023.

[9] I. T. Jolliffe, "Principal component analysis," *Springer,* 2020.

[10] A. Mahapatra, A. M. B. Nanda, A. Padhy and I. Padhy, "Concept of outlier study: The management of outlier handling with significance in Inclusive education setting," *Asian Research Journal of Mathematics,* pp. 7-25, 2020.

[11] S. S, X. L, R. L, G. F, L. S, W. Z and X. R, "An efficient density-based local outlier detection approach for scattered data," *IEEE Access,* vol. 7, 2020.

[12] Y. Roh, G. Heo and S. E. Whang, " A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Transactions on Knowledge and Data Engineering,* vol. 1, 2019.

[13] K. Phiwhorm, C. Saikaew, C. Leung, P. Polpinit and K. Saikaew, "Adaptive multiple imputations of missing values using the class center," *Big Data,* vol. 9, no. 52, 2022.

[14] M. Hasan, M. Alam, S. Roy, A. Dutta, M. Jawad and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature," *Inform. Med. Unlocked,* vol. 27, 2021.

[15] T. Boeckling, G. De Tre and A. Bronselaer, "Cleaning Data With Selection Rules," *IEEE Access,* vol. 10, 2022.

[16] V. Viallon, M. His, S. Rinaldi, M. Breeur, A. Gicquiau, B. Hemon, K. Overvad, A. Tjønneland, A. Rostgaard-Hansen and J. Rothwell, "A New Pipeline for the Normalization and Pooling of Metabolomics Data," *Metabolites,* vol. 11, 2021.

[17] B. Deng and X. Li, "A novel normalization method for imbalanced data classification based on deep learning," *Expert Systems with Applications,* vol. 16828, p. 199, 2022.

[18] V. Eric, K. Angel and G. Helena, "Measuring the effect of categorical encoders in machine learning tasks using synthetic data," pp. 92-107, 2021.

[19] A. A. Abu-Hassan and H. Al-Yaqout, "A novel parameter estimation approach for clustering imbalanced data," *Expert Systems with Applications,* vol. 113792, p. 116, 2020.

[20] R. Ketan, "Cluster quality analysis using silhouettescore," *M.S. Writing Project, Department of Computer Scienceand Electrical Engineering,* 2020.

[21] M. Rodriguez, C. Comin, D. Casanova, O. Bruno, D. Amancio, L. Costa and F. Rodrigues, "Clustering algorithms: A comparative approach," *PLoS ONE,* vol. 14, 2019.

[22] C. Patil and I. Baidari, "Estimating the optimal number of clusters k in a dataset using data depth," *Data Sci,* vol. 4, pp. 132-140, 2019.

[23] R. S. K and N. C, "Cluster quality analysis using silhouette score," 2020.

[24] T. Liu, H. Yu and R. Blair, "Stability estimation for unsupervised clustering: A review," *Comput. Stat.,* vol. 14, 2022.