JURNAL FASILKOM
Volume 13 No. 3 | Desember 2023: 349-357

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Perbandingan Metode K-Means Clustering Dan Metode Ward Dalam Mengelompokkan Pelangan Mall

Tia Iklima[1], Ardi Pujiyanta[2]
[1,2]Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1]tiaiklima1800018407@webmail.uad.ac.id, [2*]ardipujiyanta@tif.uad.ac.id

Abstract

For every company, customers are an important aspect, especially in the business world. Because customers play an important role in the progress of the company, one of which is the Mall. Therefore, it is necessary to carry out further analysis to analyze customer data. However, in analyzing customer data, it is necessary to use the right method in the process of grouping customer data. In this research, a comparison of two different methods was carried out, namely the K-means method and the Ward method. The dataset used in this research contains the behavioral patterns of each customer. To calculate the minimum distance from each data in the cluster, Euclidean Distance is used. The customer data used is Mall customer data, which is taken from public data on the Kaggle.com platform, which consists of 5 variables, namely CustomerID, Gender, Age, Total Earning, and Spending Score. The research results show that the K-means method and Ward's method can be applied to the dataset used. The grouping results obtained 4 different customer groups. The accuracy value for each method was tested using the Silhouette Coefficient method. The results of grouping customer data using the K-means method for an s(i) value of 0.67. Meanwhile, the s(i) value is 0.81 for the Ward method. Based on the research results, to determine customer groups based on the level of similarity of each object, using the Ward method is better than the KMeans method in the Clustering process on the Mall customer dataset.

Keywords: Clustering, K-Means, Ward, Silhouette Coefficient

Abstrak

Bagi setiap perusahaan pelanggan adalah aspek yang penting, apalagi dalam dunia bisnis. Karena pelanggan berperan penting dalam kemajuan perusahaan salah satunya Mall. Oleh karena itu perlu dilakukannya analisa lebih lanjut untuk menganalisis data pelanggan. Namun dalam menganalisa data pelanggan perlu menggunakan metode yang tepat dalam proses pengelompokan data pelanggan. Dalam penelitian ini dilakukan perbandingan dua metode yang berbeda, yaitu metode K-means dan metode Ward. Dataset yang digunakan dalam penelitian ini mengandung pola tingkah laku dari masing-masing pelanggan. Untuk menghitung jarak minimum dari setiap data dalam cluster digunakan Euclidean Distance. Data pelanggan yang digunakan yaitu data pelanggan Mall, yang diambil dari data public pada platform Kaggle.com, yang terdiri dari 5 variabel yaitu CustomerID, Gender, Age, Total Earning, dan Spending Score. Hasil penelitian menunjukkan bahwa metode K-means dan metode Ward dapatditerapkan pada dataset yang digunakan. Hasil pengelompokkan didapatkan 4 kelompok pelanggan yang berbeda. Nilai akurasi pada masing-masing metode dilakukan pengujian dengan menggunakan metode Silhouette Coefficient. Hasil pengelompokan data pelanggan dengan metode K-means untuk nilai s(i) sebesar 0.67.Sedangkan nilai s(i) sebesar 0.81 untuk metode Ward. Berdasarkan hasil penelitian, untuk mengetahui kelompok pelanggan berdasarkan tingkat kemiripan setiap objek, penggunaan metode Ward lebih baik dibandingkan dengan metode KMeans dalam proses Clustering pada dataset pelanggan Mall.

Kata kunci: Clustering, K-Means, Ward, Silhouette Coefficient

1. Pendahuluan

Saat ini merupakan era bisnis, banyak sekali masyarakat yang mulai membangun bisnis dan malah sudah memiliki bisnis. Baik Seseorang maupun kelompok ataupun organisasi akan melakukan aktivitas dengan tujuan untuk bisa mendapatkan keuntungan atau laba, dimana didalamnya terdapat aktivitas pembelian, produksi, penjualan, maupun pertukaran barang dan jasa[1]. Oleh karena itu, pada pusat perbelanjaan terdapat berbagai macam persaingan bisnis. Pusat perbelanjaan sendiri ialah sekelompok pengecer dan pengusaha yang merencanakan, mengembangkan, membangun, memiliki dan mengelola properti tunggal[2]. Pusat perbelanjaan yang banyak disenangi oleh masyarakat salah satunya *Mall*. *Mall* adalah suatu pusat belanja yang didalamnya meliputi suatu komplek pertokoan, dimana transaksi jual beli maupun pertukaran barang dan jasa dilakukan, serta sebagai tempat berkumpul dan hiburan[3]. *Mall* didalamnya terdapat berbagai macam kebutuhan yang dibutuhkan oleh setiap individu seperti sandang dan pangan. *Mall* memiliki daya tarik tersendiri bagi masyarakat karena *Mall* tidak hanya untuk melakukan perbelanjaan saja namun *Mall* juga sebagai tempat berkumpul dan berekreasi yang dapat dikunjungi seperti Bioskop, *Restaurant* atau *Foodcourt* maupun arena permainan, dan tidak dapat dipungkiri

P-ISSN : 2089-3353
E-ISSN : 2808-9162

juga banyak juga pelanggan yang hanya berjalan-jalan saja atau hanya melihat-lihat saja tidak membeli apapun dan tidak melakukan transaksi apapun. Karena *Mall* yang ingin berhasil harus memperhatikan hubungan antara perusahaan dengan pelanggannya. Keterkaitan antara perusahaan dan pelanggan menjadi faktor yang berpengaruh karena demi membantu perkembangan serta kelangsungan dari perusahaan dan tentunya pelanggan berperan penting dalam meningkatkan laba perusahaan. Perusahaan harus mengenali rata-rata usia pelanggan yang datang ke mall dan keterkaitannya dengan produk yang dibeli. Perusahaan juga harus dapat mengenali usia pelanggan dan besarnya pengeluaran yang dikeluarkan dalam membeli produk yang ditawarkan. Perusahaan juga harus tahu kebutuhan yang diperlukan pelanggan, sehingga pelanggan tidak meninggalkan perusahaan. Oleh karena itu perlu dilakukannya analisa lebih lanjut untuk menganalisis data pelanggan dengan penggunaaan metode yang baik dalam pengelompokkan pelanggan. Agar tujuan tersebut tercapai, akan digunakan dataset pelanggan *Mall* dalam mengelompokan objek berdasarkan tingkat kemiripan antar objek. Dengan dilakukannya analisis dataset pelanggan ini dengan penggunaan metode yang baik, perusahaan dapat mengetahui kelompok dari masing-masing pelanggannya dengan hasil akurasi yang tinggi.

Penelitian[4], meggunakan metode *K-means* dan metode *Ward* dalam penentuan kelompok pelanggan. Metode *K-means* dari sisi waktu proses komputasinya relatif singkat serta memiliki ketelitian yang cukup tinggi terhadap ukuran objek. Untuk pengolahan objek dalam jumlah besar lebih terukur dan efisien[5]. Sedangkan dalam penggunaan metode *Ward* dapat meminimumkan jumlah kuadrat (*SSE*) dalam pengelompokan dari metode *hirarki* lainnya[6]. Dalam pengujian untuk mengetahui seberapa baik penempatan objek dalam suatu *Cluster* digunakan *coefficient*.

Penelitian[7] menerapkan metode *Ward* Dan Algoritma *K-means* untuk mengelompokkan kabupaten atau kota berdasarkan kasus penyakit di Sulawesi selatan pada tahun 2017. Kedua metode tersebut kemudian dibandingkan dengan menggunakan uji *Silhouette Coefficient*.

Dalam Penelitiannya[8] menciptakan titik-titik pusat penyebaran virus corona di beberapa wilayah kecamatan Kota Cirebon yang sering dijadikan tempat beraktivitas, seperti pasar, perkantoran, dan lain-lain. Sehingga didapatkan pola *Cluster* penyebaran virus corona. Hasil yang didapatkan sebagai bahan masukan serta skala prioritas untuk pemerintah Kota Cirebon.

Dalam penelitian[9] melakukan proses pengelompokkan untuk mengetahui sejauh mana tingkat pemerataan serta karakteristik pendidikan pada kecamatan di Sulawesi Barat. Penggunaan metode ini untuk memperoleh *Cluster* yang memiliki *varian* minimum. Perhitungan jarak menggunakan *Euclidean Distance,* dan jumlah *Cluster* yang digunakan n=3, terdiri dari beberapa kecamatan pada setiap kelompoknya.

Tujuan penelitian[6] mengelompokkan obat, dimana obat-obat yang mirip akan dijadikan satu kelompok atau *Cluster* data tertentu. Untuk mengetahui tingkat kemiripan dari data maka dilakukan perhitungan jarak dengan menggunakan *Euclidean Distance*. Semakin kecil jarak antar data maka semakin tinggi tingkat kemiripan dari data tersebut. Hasil penelitian menunjukkan dengan jumlah *Cluster* n=2 merupakan *Cluster* yang ideal.

Penelitian[10] melakukan pengelompokkan provinsi-provinsi di Indonesia berdasarkan tingkat keparahan dampak *Covid-19* terhadap perekonomian pada masing- masing daerah. H asil penelitian menggunakan metode *Ward* dengan jarak *Euclidean,* diperoleh hasil terbaik dibandingkan dengan metode lain pada pengelompokkan dampak ekonomi *Covid-19* dengan nilai *Silhouette* sebesar 0.48.

Penelitian[11] mengelompokkan jenis obat yang memiliki *varian* yang banyak. Pada bidang Kesehatan obat merupakan kebutuhan sehari-hari. Oleh karena itu, dibutuhkannya pengelompokkan jenis obat dengan menggunakan dua metode yang berbeda yaitu Metode *K-means* Dengan *Hierarchical Clustering Single Linkage,* untuk menentukan metode mana yang lebih optimal.

Penelitian[12] menerapkan metode *K-Means Clustering* dan *Hierarchical Clustering,* yang digunakan untuk mengetahui metode mana yang lebih bagus dan menghasilkan tingkat kemiripan yang optimal. Cara yang dilakukan dengan membandingkan dua persamaan tersebut untuk mencari faktor yang menyebabkan persamaan dan perbedaannya.

Berdasarkan latar belakang dan kajian pustaka pada penjelasan di atas, maka penulis menindaklanjuti permasalahan tersebut dengan melakukan penelitian metode mana yang lebih baik penggunaannya dalam proses *Clustering* data pelanggan *Mall*. Dari dua metode yaitu *K-means* dan *Ward,* kita dapat mengetahui nilai akurasi dari uji validitas dengan menggunakan uji *Silhouette Coefficent* yang telah dilakukan, metode manakah yang memiliki performa lebih baik.

2. Metode Penelitian

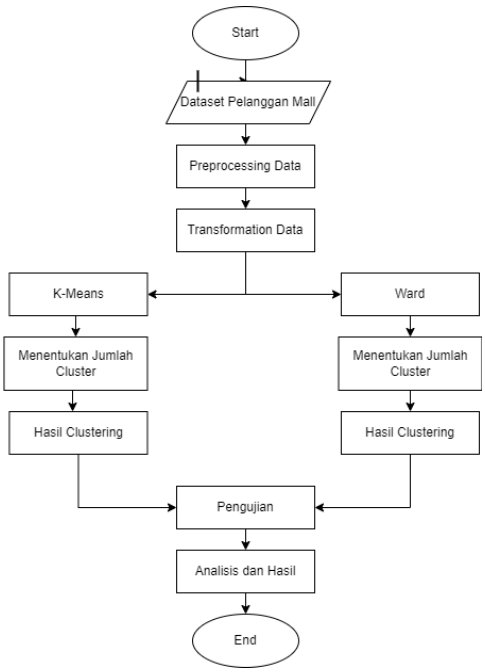
Data yang digunakan dalam penelitian ini berupa dataset yang bersifat public, yang berasal dari sebuah platform yang Bernama Kaggle.com. Gambar 1 menunjukkan langkah-langkah dari penelitian yang dilakukan. Gambar 1 menunjukkan bahwa penelitian dimulai dari dataset sekunder, analisis kebutuhan, pembuatan bagan alir, implementasi, dan yang terakhir pengujian. Pada tahap dataset sekunder yaitu kegiatan mencari data yang telah tersedia sebelumnya. Setelah data diperoleh, tahap selanjutnya yaitu analisis kebutuhan dimana dataset akan di analisis untuk

P-ISSN : 2089-3353
E-ISSN : 2808-9162

mengetahui jumlah data dan variabel yang ada pada data tersebut[13],[14].

Gambar 1. Tahapan Penelitian

Setelah analisis kebutuhan, yaitu pembuatan bagan alir (flowchart) yang memuat tahapan-tahapan proses dari analisis dataset yang dilakukan. Gambar 2 adalah flowchart analisis dataset pelanggan mall.



Gambar 2. Flowchart Analisis Dataset Pelanggan Mall

Berdasarkan Gambar 2, maka dapat diuraikan proses analisis dataset pelanggan mall dengan menggunakan metode KMeans dan Ward yaitu :

1. Dataset Pelanggan *Mall*

   Data yang digunakan yaitu dataset pelanggan *Mall*. Dimana parameter dalam dataset ini yaitu *Customer ID*, *Gender*, *Age*, *Total Earning*, dan *Spending Score.*

2. *Preprocessing* Data

   Pada tahap ini dilakukan *drop* variabel pada dataset pelanggan *Mall* yang tidak digunakan yaitu variabel *Customer ID* dan *Gender*. Sedangkan variabel yang digunakan yaitu *Age*, *Total Earning*, dan juga *Spending Score*.

3. *Transformation* Data

   Setelah dilakukan *preprocessing* maka tahap selanjutnya yaitu transformation data. Pada tahap ini dataset akan dinormalisasikan. Normalisasi dilakukan untuk menskalakan nilai data dalam rentang nilai tertentu, seperti -1 hingga 1 atau 0 hingga 1.

4. Proses *Clustering*

   Tahap selanjutnya yaitu tahap *Clustering*[15]. Dalam analisis dataset pelanggan akan dikelompokan menggunakan dua metode yaitu metode *KMeans* dan *Ward*. Metode *KMeans* dan *Ward* ini akan membagi data menjadi beberapa kelompok (*Cluster*). Dataset akan dikelompokan ke dalam kelompok dengan karakteristik yang sama dalam satu *Cluster,* oleh metode *Kmeans.* namun memiliki karakteristik yang berbeda dengan *Cluster* yang lain, berdasarkan jarak minimum masing-masing objek terhadap *Centroid*. Sedangkan metode *Ward* merupakan metode *varian*. Dimana metode ini akan menggabungkan objek berdasarkan *varian* nilai terkecil.

5. Menentukan Jumlah *Cluster*

   Dalam menentukan jumlah *Cluster* dalam menganalisis dataset pelanggan pada metode *KMeans* menggunakan metode *elbow* (metode siku)[16], sedangkan untuk menentukan jumlah *Cluster* pada metode *Ward* menggunakan *Dendogram* (melihat hasil representasi *visual* bagaimana *Cluster* terbentuk)[17]. Dengan mencari nilai sum of square error (SSE). Nilai sum of square error(*SSE*) merupakan hasil penjumlahan dari seluruh jarak masing-masing data dengan *Centroid*. Semakin kecil nilai *SSE*, semakin seragam data yang ada di dalam masing-masing *Cluster.* Teknik *SSE* digunakan untuk mengevaluasi jumlah *Cluster* yang dihasilkan dari pengujian dengan *K-means* dan *Ward.* Berikut rumus untuk mencari nilai SSE[18].

$$SSE = \sum_{k-1}^{k} \quad \sum xi \in sk \, |xi - ck|^2 \qquad \ldots\ldots 1$$

Keterangan :

Xi: Nilai atribut dari data ke-i

Ck: Nilai atribut titik pusat *Cluster* ke-I

6. Hasil *Clustering*

   Dari proses *Clustering* menggunakan dua metode berbeda, maka dihasilkan kelompok pelanggan *Mall* berdasarkan nilai k yang telah ditentukan sebelumnya pada masing-masing metode.

7. Pengujian

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Tahap selanjutnya yaitu pengujian untuk mengetahui kualitas dan kekuatan hasil *Cluster* dengan menggunakan metode *Silhouette Coefficient* untuk memvalidasi hasil *Cluster*. Nilai *Silhouette Coefficient* pada interval -1 hingga 1. Semakin dekat nilai rata-rata *s(i)* dengan 1, maka semakin baik pengelompokan data dalam satu *Cluster*.

8. Analisis dan Hasil
   Pada tahap ini, hasil *Clustering* dataset pelanggan *Mall* dilakukan analisis untuk mengetahui kelompok pelanggan berdasarkan variabel Age, *Total Earning*, dan *Spending Score*. Seperti jumlah anggota pada masing-masing *Cluster*, masing-masing *Cluster* masuk pada kelompok pelanggan dan hasil *s(i)* pada masing-masing *Cluster*.

3. Hasil dan Pembahasan

3.1. Dataset Pelanggan Mall

Penelitian ini menggunakan data *public* yang bernama *Mall* dataset1000[19]. Dataset tersebut berisi 5 variabel, dalam penelitian ini hanya 3 variabel sebanyak 1000 data, yaitu *Age, Total Earning, Spending Score*[20]. Dimana *Age* merupakan umur dari pelanggan tersebut, *Total Earning* adalah pendapatan pertahun, sedangkan *Spending Score* merupakan nilai yang diberikan oleh pihak *mall* kepada pelanggan dengan rentang 1-100 dimana variabel tersebut berdasarkan tingkah laku pembelian, sifat dan *behaviour*. Tabel 1 menunjukkan dataset pelanggan mall yang digunakan.

Tabel 1 Dataset Pelanggan Mall

CustomerID	Gender	Age	Total Earning	Spending Score
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40
5	Female	22	17	76
…	…	…	…	…
995	Female	22	84	56
996	Female	22	65	42
997	Male	23	115	19
998	Male	52	86	24
999	Male	58	124	28

3.2. Preprocessing Data

Langkah preprosesing dataset dengan cara melakukan *drop* variabel yang tidak digunakan, yaitu variabel *Customer ID* dan *Gender,* merupakan variabel yang tidak digunakan. Sedangkan variabel yang digunakan yaitu *Age*, *Total Earning*, dan juga *Spending Score,* ditunjukkan oleh Tabel 2.

Tabel 2 Dataset Pelanggan Mall

No.	Age	Total Earning	Spending Score
0	19	15	39
1	21	15	81
2	20	16	6
3	23	16	77
4	31	17	40
5	22	17	76
…	…	…	…
995	22	84	56
996	22	65	42
997	23	115	19
998	52	86	24
999	58	124	28

3.3. *Transformation* Data

Setelah dilakukan *preprocessing,* maka tahap selanjutnya yaitu transformation data. Pada tahap ini dataset akan dinormalisasikan. Gambar 3 menunjukkan hasil normalisasi data dengan metode min-max. metode min-max digunakan untuk menskalakan masing-masing variabel yang digunakan sehingga kumpulan data menjadi skala mulai dari 0 (min) hingga 1 (max). Normalisasi dataset perlu dilakukan karena dataset berjumlah besar dan variabel-variabel dari dataset memiliki skala atau rentang yang berbeda.

```
array([[0.01923077, 0.        , 0.38383838],
       [0.05769231, 0.        , 0.80808081],
       [0.03846154, 0.00740741, 0.05050505],
       ...,
       [0.09615385, 0.74074074, 0.18181818],
       [0.65384615, 0.52592593, 0.23232323],
       [0.76923077, 0.80740741, 0.27272727]])
```

Gambar 3 Normalisasi Data

3.4. Clustering

3.4.1. Metode KMeans

Pada tahap ini data pelanggan mall akan dibagi menjadi beberapa kelompok. Kelompok pelanggan ditentukan berdasarkan nilai k dengan metode elbow. Metode ini mencari nilai SSE terbaik. Gambar 4, merupakan nilai SSE untuk mencari nilai k optimal dan menunjukkan masing-masing jumlah *Cluster* 1 sampai 10. Pada *Cluster* 1, sampai 3 merupakan *Cluster* dengan nilai tertinggi, namun setiap *Cluster* nya mengalami penurunan yang cukup signifikan.

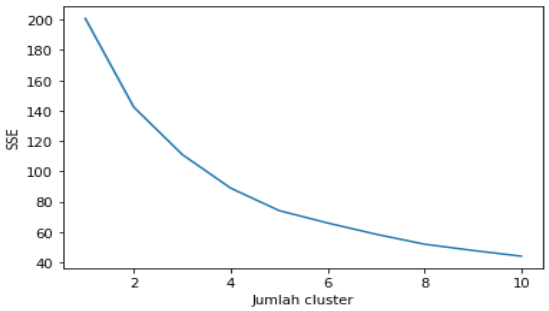
```
1 200.78049968390684
2 142.36612555068746
3 111.1046101372273
4 88.93720897271189
5 74.13052296901377
6 65.95852321360094
7 58.52797143308025
8 51.976597713070895
9 47.832120543700825
10 44.03942976370978
```

Gambar 4 Nilai SSE

Gambar 5 menunjukkan grafik elbow grafik yang digunakan untuk menentukan jumlah *Cluster* berdasarkan nilai *SSE* yang mengalami penurunan drastis. Kualitas nilai *Cluster* semakin berkurang, ketika nilai *SSE* semakin besar. Sebaliknya semakin kecil nilai *SSE*, semakin baik kualitas *Cluster*.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Grafik membentuk siku ketika jumlah *Cluster* k=4. Oleh karena itu, data pelanggan *mall* akan dibagi menjadi 4 kelompok pelanggan.



Gambar 5 Grafik Elbow

Gambar 6 menunjukkan nilai *centroid* yang digunakan dalam proses *clustering* dengan metode *KMeans*

```
array([[0.36472653, 0.68280524, 0.79421592],
       [0.45659561, 0.73125093, 0.30607278],
       [0.18914213, 0.23734187, 0.50189134],
       [0.6478022 , 0.24740741, 0.53753865]])
```

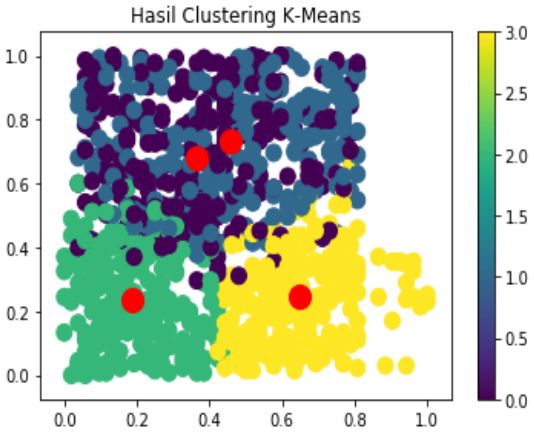
Gambar 6 Centroid KMeans

Pada Gambar 7 merupakan hasil clustering pada masing-masing pelanggan. Sedangkan pada Gambar 8 merupakan plot hasil *clustering KMeans* dimana nilai k sebanyak 4 yaitu *cluster* 0 sampai *cluster* 3. Untuk melihat seberapa banyak anggota pada setiap *cluster* dapat dilihat pada Gambar 9, dimana jumlah anggota terbanyak yaitu pada *cluster* 0 dengan jumlah anggota sebanyak 264 pelanggan. Sedangkan dalam *cluster* 1 sampai *cluster* 3 selisih jumlah anggotanya tidak terlalu jauh.

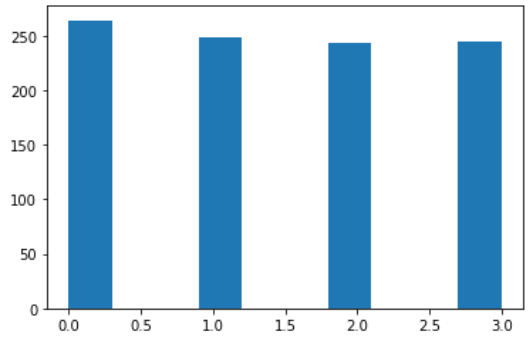
	Age	Total Earning	Spending Score	Cluster
0	0.019231	0.000000	0.383838	2
1	0.057692	0.000000	0.808081	2
2	0.038462	0.007407	0.050505	2
3	0.096154	0.007407	0.767677	2
4	0.250000	0.014815	0.393939	2
...	...	...	...	...
995	0.076923	0.511111	0.555556	2
996	0.076923	0.370370	0.414141	2
997	0.096154	0.740741	0.181818	1
998	0.653846	0.525926	0.232323	1
999	0.769231	0.807407	0.272727	1

1000 rows × 4 columns

Gambar 7 Hasil Clustering KMeans



Gambar 8 Plot Hasil Clustering KMeans



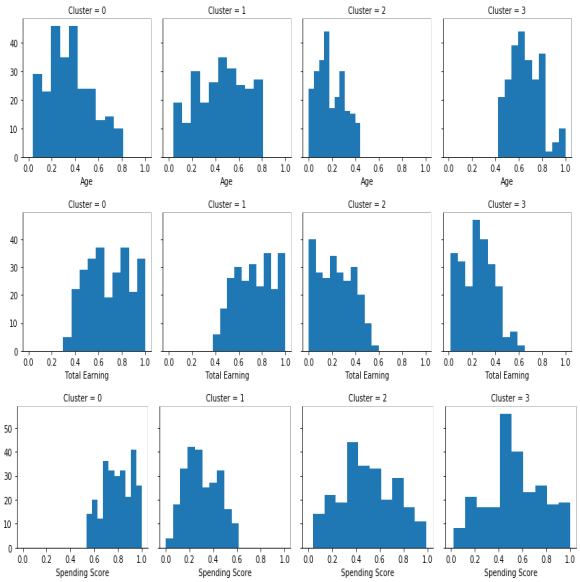
Gambar 9 Histogram Variabel Cluster

Pada gambar 10 menjelaskan *Histogram Cluster* pada masing-masing variabel, dimana :

1. Pada *Cluster* 0 merupakan kelompok pelanggan yang memiliki penghasilan dan pengeluaran yang tinggi, kelompok pelanggan ini merupakan kelompok pelanggan dengan rata-rata usia 30 sampai 80 tahun. Kelompok pelanggan jenis ini menjadi target yang menguntungkan untuk semua jenis barang.
2. Pada *Cluster* 1 merupakan kelompok pelanggan ideal dalam mempertahankan kehidupan yang optimal dalam menghasilkan dan membelanjakan. Kelompok pelanggan ini biasanya suka membeli dengan harga murah tetapi memberikan kualitas yang premium.
3. Kelompok pelanggan yang boros karena memiliki pendapatan yang sedikit namun memiliki pengeluaran yang besar dan rata-rata usia pelanggan dalam jenis ini 20 tahun, ditunjukkan pada *Cluster* 2. Pelanggan jenis ini apabila terdapat *trend* terbaru maka tidak seikit dari mereka akan melakukan pembelanjaan (tidak banyak berpikir).

P-ISSN : 2089-3353
E-ISSN : 2808-9162

4. Pada *Cluster* 3 sama dengan *Cluster* 2 termasuk pelanggan yang boros karena memiliki pendapatan sedikit namun memiliki pengeluaran yang besar, namun rata-rata usia dalam kelompok pelanggan ini usia dewasa bahkan ada juga yang lansia.



Gambar 10 Histogram Evaluasi Cluster KMeans

3.5. Metode Ward

Berikut merupakan hasil clustering pada dataset pelanggan mall dengan menggunakan metode ward, dimana k=4.

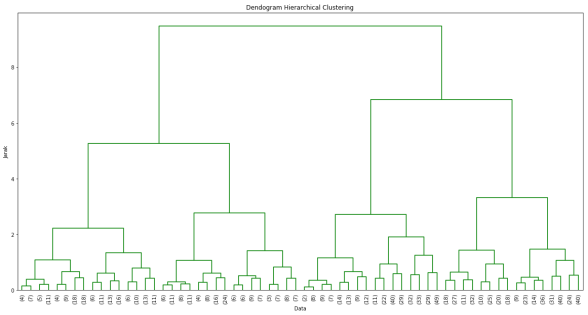
Gambar 11 merupakan hasil *clustering* setiap pelanggan dengan menggunakan metode *ward*. Dimana hasil *clustering* tersebut didapatkan dari perhitungan nilai *SSE* objek terhadap *centroid*.

	Age	Total Earning	Spending Score	Cluster
0	0.413925	0.326783	0.849635	3
1	0.247025	0.176446	0.952809	3
2	0.760286	0.608229	0.228086	1
3	0.280697	0.195267	0.939723	3
4	0.580683	0.318439	0.749269	3
...	...	...	...	...
995	0.212921	0.812971	0.541981	0
996	0.273445	0.807905	0.522031	0
997	0.193592	0.967960	0.159924	2
998	0.503268	0.832327	0.232277	2
999	0.415091	0.887436	0.200389	2

1000 rows × 4 columns

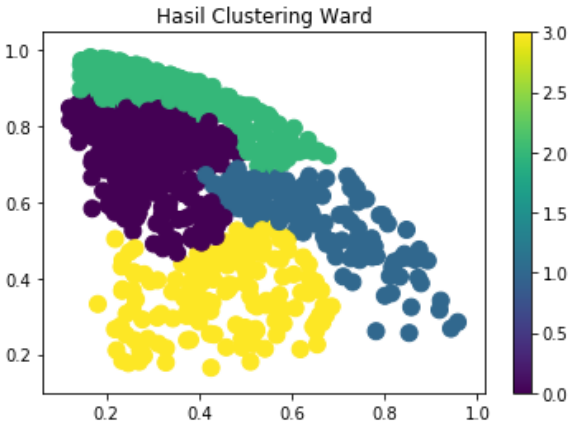
Gambar 11 Hasil Clustering Ward

Gambar 12 merupakan *Dendogram* dari hasil *Clustering*. Dimana *Dendogram* sendiri merupakan representasi *visual* bagaimana objek terbentuk *Cluster* dengan penggunaan metode *Ward*. Dari hasil *Dendogram* diatas maka penulis ingin membagi data menjadi 4 kelompok yang diambil dari garis *vertical* terpanjang pada *Dendogram* tersebut.



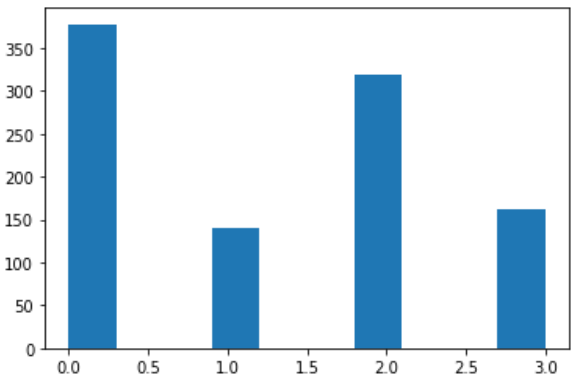
Gambar 12 Dendogram Ward

Gambar 13 menunjukkan hasil *Clustering* berdasarkan variabel *Age*, *Total Earning* dan *Spending Score*. Dari gambar tersebut dapat dilihat bahwa *Cluster* yang terbentuk dimana k=4 terjadinya *overlapping* (tumpang tindih) antar setiap objek pada penggunaan metode *Ward*.



Gambar 13 Plot Hasil Clustering Ward

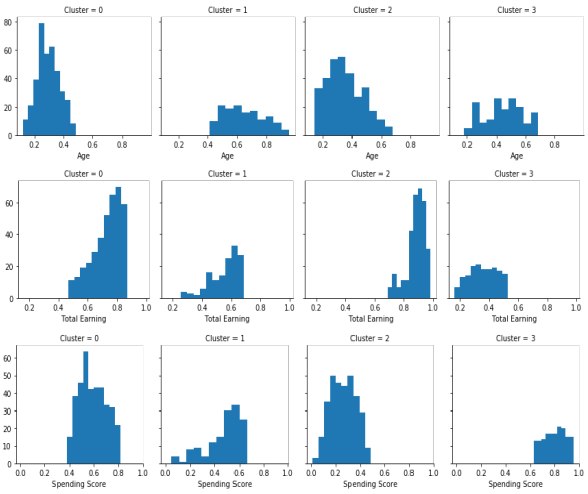
Gambar 14 merupakan *Histogram* yang menunjukan bahwa jumlah anggota pada *Cluster* 0 terbanyak yaitu sebanyak 378 pelanggan dibandingkan dengan *Cluster* yang lainnya. Oleh karena itu dapat disimpulkan *Cluster* yang memiliki anggota paling sedikit yaitu *Cluster* 1.



Gambar 14 Histogram Cluster Ward

P-ISSN : 2089-3353
E-ISSN : 2808-9162

.



Gambar 15 Histogram Evaluasi Cluster Ward

Dapat dilihat pada gambar 15 merupakan *Histogram* pada variabel *Cluster* pada masing-masing variabel, dimana :

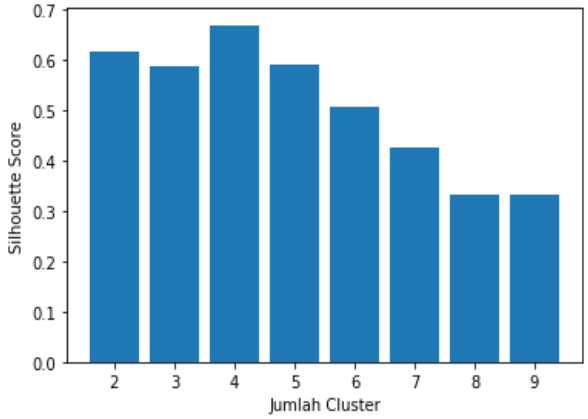
1. Pada *Cluster* 0 merupakan kelompok pelanggan yang memiliki penghasilan dan pengeluaran yang tinggi, kelompok pelanggan ini merupakan kelompok pelanggan dengan rata-rata usia 20 sampai 40 tahun. Kelompok pelanggan jenis ini menjadi target yang menguntungkan untuk semua jenis barang.
2. Pada *Cluster* 1 merupakan kelompok pelanggan ideal dalam mempertahankan kehidupan yang optimal dalam menghasilkan dan membelanjakan. Kelompok pelanggan ini biasanya suka membeli dengan harga murah tetapi memberikan kualitas yang premium
3. Pada *Cluster* 2 merupakan kelompok pelanggan yang sukamenabung karena tidak suka dengan pengeluaran yang tidak perlu (berpikir dalam belanja).
4. Pada *Cluster* 3 merupakan kelompok pelanggan yang boros karena memiliki pendapatan yang sedikit namun memiliki pengeluaran yang besar

3.6. Pengujian

3.6.1 Pengujian KMeans

Pengujian pada metode KMeans menggunakan silhouette coefficient untuk mengevaluasi cluster. Berikut merupakan nilai silhouette pada masing-masing cluster dengan pada metode Kmeans.

Gambar 16 merupakan *histogram* dari jumlah *Cluster* 2 sampai 9 berdasarkan nilai *silhouette*. Dimana nilai *silhouette* pada k=4 merupakannilai tertinggi. Dan dari hasil perhitungan nilai *Silhouette Score* yang didapatkan bahwa k=4 merupakan jumlah *Cluster* yang optimal karena nilai *s(i)* mendekati 1. Hal tersebut menunjukkan bahwa pengelompokandata pada jumlah *Cluster* k=4 semakin baik.



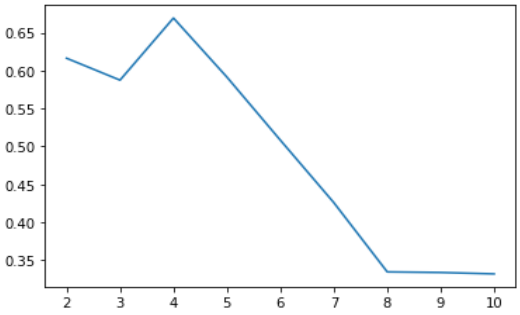
Gambar 16 Silhouette Coefficient KMeans

Gambar 17 menunjukkan nilai *score silhouette* pada jumlah *Cluster* 2 sampai 10. Dan dari hasil nilai *silhouette* tersebut didapatkan nilai hasil akurasi tertinggi yaitu pada k=4 sebanyak 0.67. Berdasarkan pengelompokan data dapat disimpulkan bahwa masing-masing *Cluster* baik karena nilai *s(i)* mendekati 1. Nilai *silhouette* tersebut dalam kriteria pengukuran pengelompokan berdasarkan *Silhouette Coefficient* termasuk dalam kriteria yang baik karena masuk dalam rentang $0.5 < s(i) \leq 0.7$.

```
Jumlah klaster = 2 , nilai average_silhoutte = 0.6163812903236323
Jumlah klaster = 3 , nilai average_silhoutte = 0.5874580035819249
Jumlah klaster = 4 , nilai average_silhoutte = 0.6694374290373569
Jumlah klaster = 5 , nilai average_silhoutte = 0.5909705752089849
Jumlah klaster = 6 , nilai average_silhoutte = 0.5082384995669073
Jumlah klaster = 7 , nilai average_silhoutte = 0.42585620603470226
Jumlah klaster = 8 , nilai average_silhoutte = 0.4228517213369361
Jumlah klaster = 9 , nilai average_silhoutte = 0.3342927370365971
Jumlah klaster = 10 , nilai average_silhoutte = 0.33137599056444667
```

Gambar 17 Nilai Silhouette Coefficient

Gambar 18 merupakan grafik nilai *s(i)*, dimana nilai *silhouette* pada k=2 dan k=3 nilai *silhouette* tidak teratur dan pada nilai *silhouette* k=4 merupakan nilai tertinggi (maksimum) dan apabila terjadinya penambahan jumlah *Cluster* maka nilai *silhouette* mulai melambat menurun yang menunjukkan bahwa dengan peningkatan *Cluster* baru tidak banyak perbaikan yang signifikan pada metode ini.

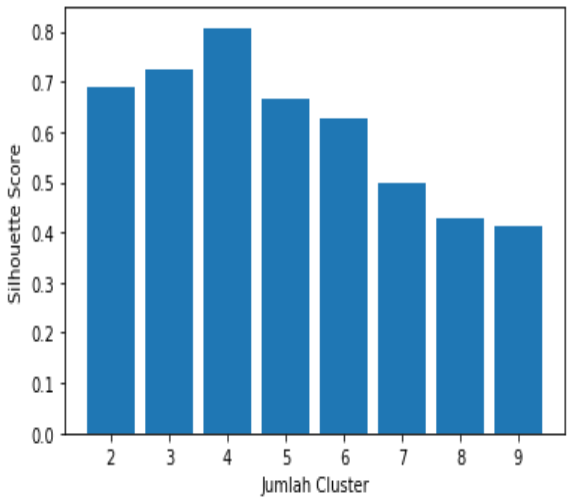


Gambar 18 Grafik Hasil Silhouette KMeans

3.6.2 Pengujian Ward

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Gambar 19 merupakan *histogram* dari nilai *Silhouette Score*, dimana nilai *s(i)* tertinggi pada k=4 yaitu menyentuh nilai 0.8 lebih. Dapat disimpulkan bahwa kualitas *Cluster* pada k=4 memiliki kekuatan dan kualitas hasil *Cluster* yang baik.



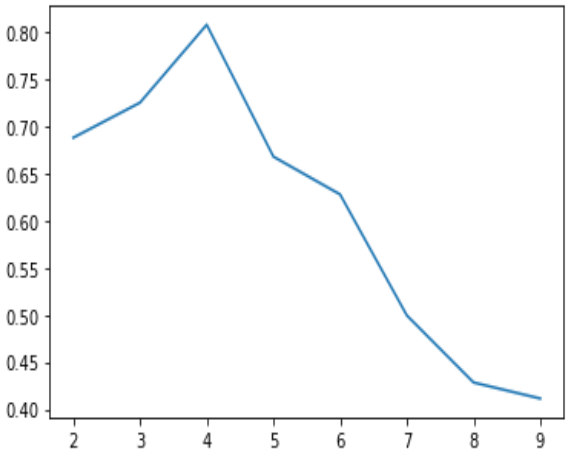
Gambar 19 Silhouette Coefficient Ward

Gambar 20 merupakan nilai *Silhouette Score* menggunakan metode *Ward*. Dimana nilai *s(i)* tertinggi pada k=4 dengan nilai *s(i)* sebesar 0.81. Dapat disimpulkan hasil pengelompokan objek terhadap masing-masing *Cluster* semakin baik karena nilai *s(i)* mendekati 1. Nilai *silhouette* tersebut dalam kriteria pengukuran pengelompokan berdasarkan *Silhouette Coefficient* termasuk dalam kriteria yang kuat karena masuk dalam rentang $0.7 < s(i) \leq 1.0$.

```
[0.6876455969930337,
 0.7247365408641112,
 0.8072843308555466,
 0.6675580381049554,
 0.6275132544349543,
 0.49941764631527247,
 0.4283913741126536,
 0.41138382649013716]
```

Gambar 20 Nilai Silhouette Ward

Dapat dilihat grafik pada Gambar 21 bahwa nilai *Silhouette Scores* dengan jumlah *Cluster* 2 sampai 4 terjadinya peningkatan dan nilai *silhouette* tertinggi dimiliki oleh k=4. Tetapi jika dilakukan penambahan jumlah *Cluster* maka nilai *Silhouette Score* akan semakin turun. Dari grafik tersebut disimpulkan bahwa jika terjadinya penambahan jumlah *Cluster* maka tidak ada perbaikan secara signifikan pada penggunaan metode ini.



Gambar 21 Grafik Nilai Silhouette Ward

Dari hasil pengujian pada masing-masing metode menunjukkan pengukuran nilai *silhouette* untuk setiap sampel berkisarantara -1 sampai 1. Jika rata-rata *s(i)* mendekati angka 1 berarti semakin baik pengelompokkan data di dalam satu *Cluster*. Sedangkan 0 tidak adakesamaan didalam *Cluster* dan tidak ada perbedaan juga pada *Cluster* yang lain. Dan jika nilai *s(i)* mendekati nilai -1 pengelompokan data semakin tidak baik dalam satu *Cluster*[7].

4. Kesimpulan

Dari dataset pelanggan *Mall* tersebut didapatkan 4 kelompok pelanggan dimana kelompok pelanggan pada *Cluster* 0 rata-rata usia pelanggan ini 30 sampai 80 tahun dengan penghasilan dan pengeluaran yang tinggi, kelompok jenis ini merupakan target yang menguntungkan untuk semua jenis barang. Kelompok pelanggan pada *Cluster* 1 merupakan jenis pelanggan yang ideal karena penghasilan dan pengeluaran seimbang, pelanggan jenis ini biasanya suka membeli dengan harga murah tetapi memberikan kualitas yang *premium*. Pada *Cluster* 2 merupakan kelompok pelanggan yang boros karena memiliki pendapatan yang sedikit namun memiliki pengeluaran yang besar dan rata-rata usia pada jenis pelanggan ini 20 tahun. Pada *Cluster* 3 sama dengan kelompok pada *Cluster* 2 memiliki pendapatan sedikit namun pengeluarannya besar, yang membedakannya usia rata-rata pada kelompok pelanggan ini, yaitu usia dewasa bahkan ada juga yang lansia. Nilai *silhouette* yang didapatkan pada pengelompokan dengan metode *KMeans* sebesar 0.67 dengan struktur baik. Sedangkan pengelompokan pelanggan dengan menggunakan metode *Ward* menghasilkan nilai *silhouette* 0.81 dengan struktur kuat. Sehingga dari nilai *silhouette* dapat disimpulkan bahwa penggunaan metode *Ward* lebih baik dari penggunaan metode *KMeans* untuk proses pengelompokkan pelanggan *Mall* karena memiliki nilai *silhouette* yang lebih tinggi, yang menunjukan bahwa pengelompokan pada masing-masing *Cluster*-nya memiliki tingkat kemiripan yang tinggi.

Daftar Rujukan

[1] P. Guru, "Pengertian Bisnis." https://pendidikan.co.id/pengertian-bisnis/ (accessed Sep. 17, 2023).

[2] Wikipedia, "Pusat perbelanjaan - Wikipedia bahasa Indonesia, ensiklopedia bebas." https://id.wikipedia.org/wiki/Pusat_perbelanjaan (accessed Sep. 17, 2023).

[3] Aska, "Klasifikasi Jenis Mall dan Pusat Perbelanjaan," *Arsitur Studio*. https://www.arsitur.com/2017/12/klasifikasi-jenis-mall-dan-pusat.html (accessed Sep. 17, 2023).

[4] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.

[5] N. Putu, E. Merliana, and A. J. Santoso, "ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING," pp. 978–979, 2015.

[6] M. Paramadina, S. Sudarmin, and M. K. Aidid, "Perbandingan Analisis Cluster Metode Average Linkage dan Metode Ward (Kasus: IPM Provinsi Sulawesi Selatan)," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 1, no. 2, p. 22, 2019, doi: 10.35580/variansiunm9357.

[7] N. Afdhaliah, "Perbandingan Kinerja Algoritma Ward Dan Algoritma K-Means Dengan Uji Silhouette Coefficient," Universitas Hasanuddin, 2020.

[8] H. Gunawan and V. Purwayoga, "Data Mining Menggunakan Algoritma K-Means Clustering Untuk Mengetahui Potensi Penyebaran Virus Corona Di Kota Cirebon," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, pp. 1–8, 2022, doi: 10.32736/sisfokom.v11i1.1316.

[9] H. Hikmah, F. Fardinah, L. Qadrini, and E. Tande, "Analisis Klaster Pengelompokan Kecamatan di Sulawesi Barat Berdasarkan Indikator Pendidikan," *Saintifik*, vol. 8, no. 2, pp. 188–196, 2022, doi: 10.31605/saintifik.v8i2.383.

[10] I. N. Hasanah and A. Sofro, "Analisis Cluster Berdasarkan Dampak Ekonomi Di Indonesia Akibat Pandemi Covid-19," *MATHunesa J. Ilm. Mat.*, vol. 10, no. 2, pp. 239–248, 2022, doi: 10.26740/mathunesa.v10n2.p239-248.

[11] R. D. Firdaus, T. G. Laksana, and R. D. Ramadhani, "Pengelompokan Data Persediaan Obat Menggunakan Perbandingan Metode K-Means Dengan Hierarchical Clustering Single LinkageFirdaus, Rahmatika Diana Laksana, Tri Ginanjar Ramadhani, Rima Dias," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 2, no. 1, pp. 33–48, 2019.

[12] N. K. Zuhal, "Study Comparison K-Means Clustering dengan Algoritma Hierarchical Clustering," *Pros. Semin. Nas. Teknol. dan Sains*, vol. 1, pp. 200–205, 2022, [Online]. Available: https://jurnal.dharmawangsa.ac.id/index.php/djtechno/article/view/966/867.

[13] W. T. Saputro, M. Murhadi, and H. M. Jumasa, "Menemukan Pola Sebaran Vaksinasi Data Covid-19 di Indonesia Menggunakan Algoritma K-Means," *J. Fasilkom*, vol. 13, no. 02, pp. 244–250, 2023, doi: 10.37859/jf.v13i02.5551.

[14] D. Anisa, W. S. Ningrum, R. Kusumo, and W. Putri, "Sistem Pendukung Keputusan Penerimaan Beasiswa Menggunakan Metode Weighted Product," *TIN Terap. Inform. Nusant.*, vol. 2, no. 8, pp. 483–491, 2022, doi: 10.47065/tin.v2i8.1064.

[15] M. Benri, H. Metisen, and S. Latipa, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," *J. Media Infotama*, vol. 11, no. 2, pp. 110–118, 2015, [Online]. Available: https://core.ac.uk/download/pdf/287160954.pdf.

[16] R. T. S. Muhammad Hariyanto, "Clustering pada Data Mining untuk Mengetahui Potensi Penyebaran Penyakit DBD Menggunakan Metode Algoritma K-Means dan Metode Perhitungan Jarak Euclidean Distance," *Sist. Komput. dan Tek. Inform.*, vol. 1, no. 1, pp. 117–122, 2018.

[17] P. S. Matematika, J. P. Matematika, F. Matematika, D. A. N. Ilmu, P. Alam, and U. N. Yogyakarta, "Analisis Cluster dengan Average Linkage Method dan Ward ' s Method untuk Data Responden Nasabah Asuransi Jiwa Unit Link," 2014.

[18] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 918, 2019, doi: 10.30645/senaris.v1i0.100.

[19] Kaggle, "Dataset Mall 1000," *Kaggle*, 2023. https://www.kaggle.com/datasets/manojtolani/mall-dataset1000/ (accessed Apr. 12, 2023).

[20] R. Maulana, D. Adi Putra Pratama, N. Nugraha, and A. Rahmasari, "Implementasi Algoritma Hierarchical Clustering untuk Klasterisasi Data Pelanggan Mall (Implementation of Hierarchical Clustering Algorithm for Mall Customer Data Clustering)," *Gunung Djati Conf. Ser.*, vol. 3, pp. 0–4, 2021.