

# Segmentasi Karakteristik Pelanggan dengan K-Means Clustering dan Davies-Bouldin Index

Rachmad Fitriyanto<sup>1</sup>, Nurindah<sup>2</sup>

Program Studi Sistem Informasi  
STMik PPKIA Tarakanita Rahmawati  
Tarakan, Indonesia

e-mail: <sup>1</sup>rachmad@ppkia.ac.id, <sup>2</sup>Nur291030@gmail.com

## Abstrak

Pemahaman karakteristik pelanggan merupakan cara agar pemilik usaha mampu mengembangkan produk dan layanannya agar dapat meningkatkan keuntungan. Karakteristik pelanggan yang beraneka ragam merupakan sumber data berharga yang dapat diolah untuk merumuskan strategi usaha. Karakteristik pelanggan sebagai atribut data dapat diklusterisasikan untuk memudahkan menggali Informasi. K-Means Clustering (KMC) merupakan algoritma klusterisasi yang mudah diterapkan, namun memiliki celah dalam hal jumlah kluster yang tepat untuk mengelompokkan pelanggan. Davies-Bouldin Index (DBI) merupakan indikator yang dapat digunakan untuk mengevaluasi sebuah nilai jumlah kluster. Penelitian ini bertujuan untuk mensegmentasikan karakteristik pelanggan menggunakan KMC yang dievaluasi melalui nilai DBI. Penelitian dilakukan dalam 5 tahap, perancangan kuesioner, pengumpulan data, klusterisasi, evaluasi jumlah kluster dan analisis kluster. Data yang terkumpul dari 100 responden diklusterisasikan untuk 5 variasi jumlah kluster, 2 sampai 6 kluster. Setiap hasil klusterisasi dari 5 variasi jumlah kluster dihitung nilai DBI. Kluster dengan nilai DBI terbaik diproses untuk ekstraksi Informasi karakteristik pelanggan. Hasil penelitian menunjukkan jumlah kluster sebanyak 3 memiliki nilai DBI terbaik. Analisa karakteristik dari hasil klusterisasi tersebut menunjukkan

**Kata kunci:** Segmentasi Pelanggan, K-Means Clustering, Davies-Bouldin Index, Data Mining.

## Abstract

Understanding customer characteristics is a way for business owners to be able to develop their products and services to increase profits. Various customer characteristics are a valuable source of data that can be processed to formulate business strategies. Customer characteristics as data attributes can be clustered to make it easier to find information. K-Means Clustering (KMC) is a clustering algorithm that is easy to implement but has a gap in terms of the right number of clusters to group customers. The Davies-Bouldin Index (DBI) is an indicator that can be used to evaluate a cluster number value. This study aims to segment customer characteristics using KMC which is evaluated through the DBI value. The research was conducted in 5 stages, designing questionnaires, collecting data, clustering, evaluating the number of clusters, and analyzing clusters. Data collected from 100 respondents were clustered for 5 variations in the number of clusters, 2 to 6 clusters. Each clustering result of 5 variations in the number of clusters is calculated by the DBI value. The cluster with the best DBI value is processed for extraction of customer characteristic information. The results showed that the number of clusters of 3 had the best DBI value. The characteristic analysis of the clustering results shows.

**Keywords:** Customer Segmentation, K-Means Clustering, Davies-Bouldin Index, Data Mining.

## 1. Pendahuluan

Pelanggan atau konsumen merupakan target pemasaran yang menentukan kesuksesan proses bisnis suatu perusahaan. Pemahaman tentang karakteristik pelanggan bagi pengelola usaha merupakan salah satu cara untuk merumuskan strategi untuk pengembangan usaha[1]. Setiap individu pelanggan memiliki karakteristik yang bisa sama dan bisa juga berbeda dengan individu lainnya. Semakin banyak pelanggan maka semakin bervariasi karakteristik yang diperoleh. Usaha yang dibutuhkan dalam proses analisa untuk menggali Informasi tentang karakteristik pelanggan bersifat linier terhadap jumlah pelanggan, semakin banyak maka semakin besar usaha yang diperlukan. Hal ini dapat diatasi menggunakan teknik data mining. Karakteristik pelanggan pada dasarnya merupakan kumpulan data-data yang dapat diproses lebih lanjut. Data yang berjumlah besar dapat dikelompokkan atau disegmentasi menjadi beberapa kelompok atau segmen yang berukuran lebih kecil, sehingga memudahkan untuk proses analisa. Proses segmentasi di dalam data mining termasuk ke dalam jenis unsupervised classification yang juga disebut dengan fungsi clustering. Clustering atau proses klusterisasi bekerja dengan cara membagi sebuah himpunan data menjadi

beberapa subhimpunan yang memiliki kesamaan karakteristik [2][3][4]. Teknik-teknik klusterisasi dalam data mining dapat dikelompokkan menjadi 2 kelompok utama, yaitu hierarchical dan partitional.

K-Means Clustering (KMC) merupakan algoritma klusterisasi yang paling populer dikarenakan kemudahannya dan kesederhanaan prosesnya[5][6]. Gambar 1 menunjukkan proses yang terjadi di dalam algoritma KMC.



Gambar 1. Alur kerja KMC[2], [7]

Penentuan jumlah kluster di awal algoritma KMC umumnya dilakukan dengan 3 cara, sesuai keinginan pengguna, trial-and-error dan teknik evolution[1][3][6][9]. Cara yang sama juga dilakukan untuk menentukan titik pusat atau centroid dari setiap kluster di tahap kedua. Pada tahap ketiga, penghitungan jarak pada KMC dapat menggunakan beberapa salah satunya dengan formula Euclidean Distance yang ditunjukkan pada persamaan 1.

$$D_{(x,y)} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

Penentuan kluster pada data yang telah dihitung jaraknya ke masing-masing centroid dilakukan dengan membandingkan jarak data ke setiap centroid yang ada. Jika data lebih dekat dengan centroid di kluster 1 maka data tersebut dikelompokkan ke kluster 1. KMC dieksekusi secara iterative selama kelompok atau nilai kluster data di iterasi berjalan sama dengan iterasi sebelumnya[1][4][7].

Meskipun KMC termasuk algoritma klusterisasi yang populer dikarenakan kemudahannya dan kesederhanaannya, namun ada beberapa kelemahan terkait cara kerjanya, salah satunya adalah penentuan jumlah kluster (K)[1][7][8]. Penentuan nilai K umumnya dilakukan sesuai keinginan pengguna seperti yang ditemui pada penelitian [11][12]. Kedua penelitian tersebut menetapkan jumlah kluster tanpa memberikan alasan pemilihan nilai tersebut, sehingga dimungkinkan terbentuknya kluster-kluster yang tidak memiliki karakteristik khusus atau kluster yang sulit dibedakan dengan kluster lainnya. Permasalahan ini dapat diatasi menggunakan Davies-Bouldin Index (DBI). DBI adalah sebuah nilai yang dihitung dari nilai jumlah kluster sebagai dasar evaluasi setelah proses klustersasi selesai. Evaluasi menggunakan nilai DBI termasuk ke dalam jenis evaluasi internal kluster, yang dilihat dari kuantitas dan kedekatan antar data hasil klusterisasi[6]. Jumlah kluster dikatakan baik apabila memiliki nilai DBI yang paling kecil. Hasil klusterisasi dikatakan menggunakan jumlah kluster yang tepat apabila jarak antar kluster bernilai maksimal, artinya area kluster satu dengan lainnya dapat dibedakan dengan jelas. Selain itu, klusterisasi dikatakan baik apabila jarak antar data di dalam sebuah kluster bernilai minimum[3]. Penelitian yang menerapkan DBI untuk klusterisasi data yang telah disebutkan sebelumnya umumnya selesai sampai dengan pemberian label kluster saja. Untuk kasus klusterisasi karakteristik pelanggan, perlu dilakukan analisis lebih lanjut untuk setiap kluster sehingga dapat diketahui atribut yang mewakili setiap kluster yang terbentuk. Penelitian ini bertujuan untuk mengkombinasikan KMC dan DBI agar menghasilkan kluster yang tepat dan menambahkan teknik analisis kluster sehingga diketahui karakteristik pelanggan per klusternya.

**2. Metode Penelitian**

Penelitian ini dilaksanakan dalam 6 tahapan utama seperti yang diilustrasikan pada gambar 2. Tahap pertama adalah penyusunan kuesioner yang dilakukan berdasarkan teori marketing mix. Kuesioner yang berhasil disusun, kemudian disebar ke 100 responden berdasarkan rata-rata pelanggan setiap harinya untuk pengumpulan data pada tahap kedua. Pengumpulan data dilakukan menggunakan kuesioner fisik dan memanfaatkan google form.



Gambar 2. Tahapan Penelitian

Tahap ketiga adalah klusterisasi menggunakan K-Means Clustering. Klusterisasi dilakukan untuk beberapa nilai K, mulai K=2,3,4 sampai dengan K=10. Setiap hasil klusterisasi dari beberapa nilai K,

kemudian dihitung nilai DBI pada tahap keempat. Setiap nilai DBI kemudian dibandingkan dan dipilih nilai DBI terbaik pada tahap kelima, yaitu nilai DBI yang paling kecil atau mendekati nol. Hasil klasterisasi yang memiliki nilai DBI terbaik kemudian dianalisis di tahap keenam untuk menghasilkan karakteristik pelanggan.

Pada tahap pertama, penyusunan kuesioner dihasilkan 16 pertanyaan tentang pelanggan seperti dirumuskan pada tabel 1.

Tabel 1. Daftar Pertanyaan

No.	Pertanyaan	No.	Pertanyaan
1	Usia Pelanggan	9	Pekerjaan
2	Jam Pembelian	10	Jarak tempat tinggal ke counter penjualan
3	Jumlah pembelian	11	Jenis minuman favorit
4	Ukuran cup yang sering dibeli	12	Kualitas desain cup minuman
5	Persepsi harga jual minuman	13	Kesesuaian harga dengan rasa minuman
6	Pengalaman mendapatkan promo penjualan	14	Fasilitas yang harus ditingkatkan
7	Kualitas kemasan untuk pembelian > 2	15	Sumber informasi tentang produk
8	Frekuensi pembelian dalam 1 pekan	16	Kepuasan atas layanan penjual

### 3. Hasil dan Pembahasan

Pembahasan hasil penelitian disusun dalam 4 bagian. Bagian pertama menunjukkan hasil pengumpulan data. Bagian kedua menjelaskan tentang hasil klasterisasi yang telah dilakukan. Bagian ketiga menjelaskan hasil penghitungan nilai DBI dan bagian keempat menjelaskan hasil analisa dari klaster yang dipilih.

#### 3.1. Pengumpulan Data

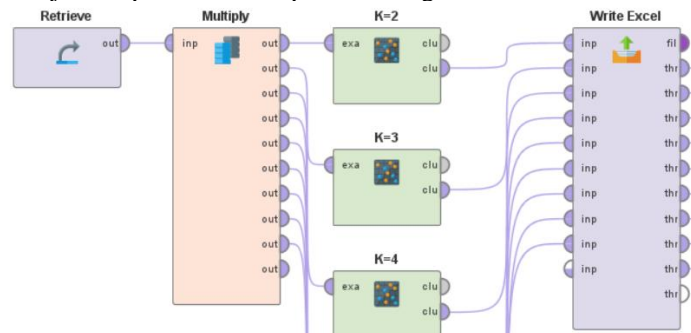
Sampel data yang terkumpul dari 100 responden ditunjukkan pada tabel 2[13].

Tabel 2. Rekapitulasi Kuesioner

R	Pertanyaan															
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
R0001	3	2	3	2	1	1	2	3	3	1	1	1	1	2	1	1
R0002	2	2	2	2	2	1	2	3	2	1	1	2	1	3	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
R0099	2	4	3	1	2	2	2	2	3	2	2	1	1	3	3	1
R0100	2	3	2	2	2	1	2	2	3	2	2	2	1	1	2	1

#### 3.2. Klasterisasi

Klasterisasi dilakukan dengan beberapa nilai K, mulai K=2,3,4 sampai K=10 menggunakan tools Rapidminer seperti ditunjukkan pada model Rapidminer di gambar 3.



Gambar 3. Model Rapidminer

Dataset yang digunakan pada model di gambar 3, digandakan terlebih dahulu menggunakan operator Multiply agar dapat diproses pada 9 operator KMC. Hasil klasterisasi dari 9 operator KMC dikonversi ke dalam sebuah file berekstensi \*.xls menggunakan operator write to excel. Contoh Hasil klasterisasi menggunakan K-Means Clustering ditunjukkan pada tabel 3.

Tabel 3. Hasil Klasterisasi dengan 9 Variasi Nilai K

Responden	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
-----------	-----	-----	-----	-----	-----	-----	-----	-----	------

R001	Klaster 2	Klaster 2	Klaster 2	Klaster 3	Klaster 4	Klaster 7	Klaster 3	Klaster 3	Klaster 1
R002	Klaster 1	Klaster 2	Klaster 1	Klaster 1	Klaster 4	Klaster 7	Klaster 3	Klaster 2	Klaster 8
...	...	...	...	...	...	...	...	...	...
R099	Klaster 2	Klaster 1	Klaster 4	Klaster 2	Klaster 5	Klaster 5	Klaster 7	Klaster 3	Klaster 4
R100	Klaster 2	Klaster 2	Klaster 1	Klaster 1	Klaster 5	Klaster 7	Klaster 5	Klaster 3	Klaster 5

**3.3. Penghitungan Nilai DBI**

Penghitungan nilai DBI dilakukan untuk hasil klasterisasi di semua variasi nilai K. Hasil penghitungan ditunjukkan pada tabel 4.

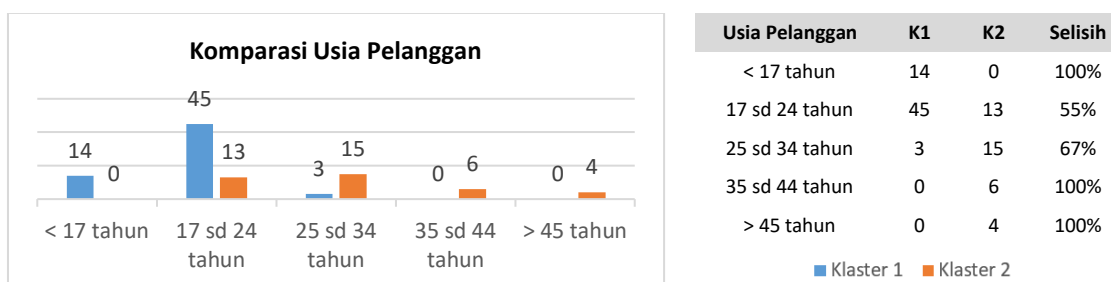
Tabel 4. Nilai DBI Di Setiap Hasil Klasterisasi

Jumlah Klaster	2	3	4	5	6	7	8	9	10
DBI	1.626	8.747	2.661	2.773	2.285	1.890	1.894	1.722	1.818

Nilai DBI terbaik pada tabel 4 adalah 1.626, hasil dari klasterisasi dengan jumlah klaster 2. Hasil klasterisasi tersebut dianalisis pada tahap selanjutnya

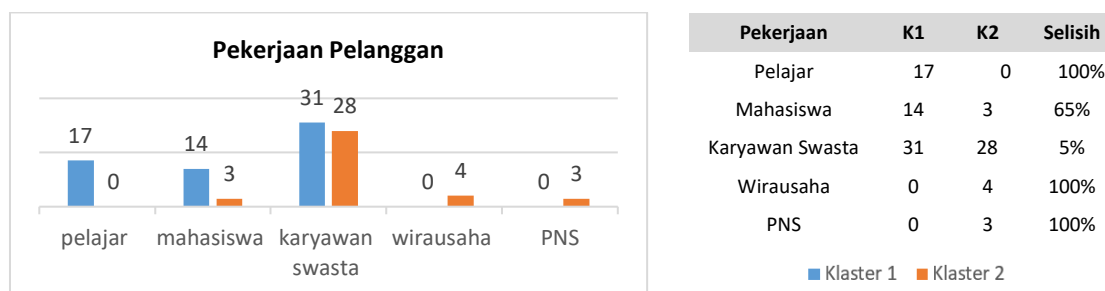
**3.4. Analisis Klaster**

Hasil klasterisasi antara klaster 1 dan klaster 2 dari perspektif usia pelanggan ditunjukkan pada gambar 3.



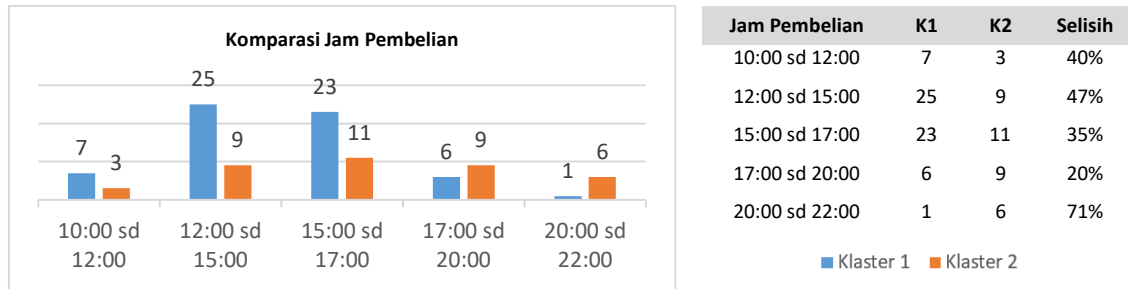
Gambar 3. Komparasi Jenis Pekerjaan Pelanggan

Klaster 1 dari perspektif usia pelanggan merepresentasikan pelanggan yang berusia di bawah 24 tahun, sedangkan dari klaster 2 merepresentasikan pelanggan yang berusia mulai 24 tahun sampai lebih dari 45 tahun. Berdasarkan usia pelanggan, klaster 1 dan klaster 2 dapat dianalisa lebih lanjut terkait dengan pekerjaan, dikarenakan faktor usia sangat mempengaruhi jenis pekerjaan. Grafik yang ditunjukkan pada gambar 4 menunjukkan perbandingan yang jelas tentang jenis pekerjaan di antara kedua klaster.



Gambar 4. Komparasi Jenis Pekerjaan Pelanggan

Karakteristik pekerjaan pelanggan yang diwakili klaster 1 adalah pelajar dan mahasiswa dengan masing-masing memiliki selisih antar klaster sebesar 100% untuk kategori pelajar dan 65% untuk kategori mahasiswa. Pada klaster 2, pekerjaan pelanggan didominasi kategori wirausaha dan PNS. Kategori karyawan swasta cenderung seimbang di kedua klaster. Segmentasi jenis pekerjaan ini bersesuaian dengan segmentasi usia pelanggan. Usia di bawah 24 tahun umumnya memiliki jenis pekerjaan sebagai siswa dan mahasiswa, sementara usia di atas 24 tahun umumnya didominasi kalangan pekerja. Karakteristik berikutnya yang terlihat jelas dari kedua klaster ada pada jam pembelian. Seperti terlihat pada gambar 5.



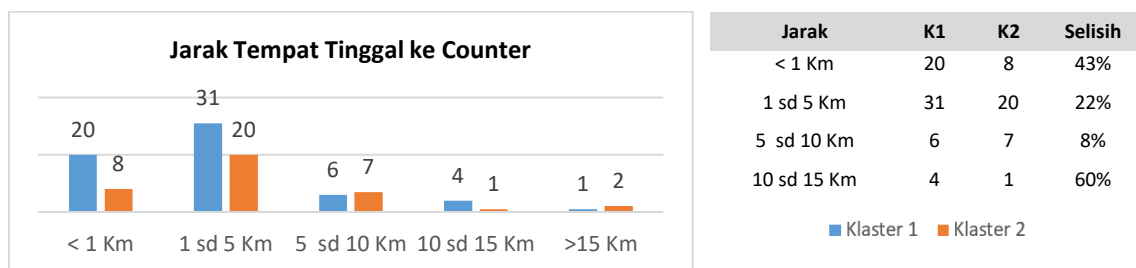
Gambar 5. Komparasi Jam Pembelian

Klaster 1 beranggotakan pelanggan yang membeli pada rentang jam 10 pagi sampai dengan jam 3 siang. Klaster 2 mewakili pelanggan yang membeli pada rentang jam 3 siang sampai jam 10 malam. Berdasarkan selisih jumlah anggota untuk setiap jam pembelian di kedua klaster, diketahui jam pembelian tersebut sesuai dengan karakteristik jenis pekerjaan pelanggan di kedua klaster. Pada analisis sebelumnya, diketahui klaster 1 mewakili pelanggan dari kelompok siswa dan mahasiswa. Kelompok ini umumnya memiliki aktifitas keseharian mulai pagi sampai dengan sore, sehingga pembelian di klaster 1 banyak terjadi di rentang jam 10 pagi sampai dengan jam 3 siang. Untuk klaster 2 pada hasil analisis sebelumnya, beranggotakan pelanggan yang bekerja, sehingga mereka hanya memiliki waktu untuk bertransaksi sepulang kerja diantara jam 5 sore sampai dengan jam 10 malam. Faktor pelanggan yang telah bekerja juga mempengaruhi jumlah pembelian seperti ditunjukkan pada gambar 6.



Gambar 6. Komparasi Jumlah Pembelian

Klaster 2 didominasi oleh pelanggan yang membeli lebih dari 2 cup, sedangkan pada klaster 1, pelanggan cenderung didominasi pembelian sejumlah 1 dan 2 cup saja. Jika analisis ditarik dari perspektif usia dan jam pembelian, hal ini sesuai dengan jumlah pembelian. Klaster 2 beranggotakan pembeli dengan usia di atas 24 tahun yang merupakan usia pekerja. Faktor usia dan status pekerjaan ini sejalan dengan jam pembelian di klaster 2 yang terjadi pada rentang jam 3 sore sampai dengan jam 10 malam. Karakteristik terakhir yang dapat digunakan untuk mendeskripsikan pelanggan adalah jarak tempat tinggal pelanggan ke counter seperti ditunjukkan pada grafik di gambar 7.



Gambar 7. Komparasi Jarak Tempat Tinggal

Pada faktor jarak tempat tinggal ke counter, pada klaster 1 didominasi oleh pelanggan yang bertempat tinggal dibawah 5 km, sedangkan klaster 2 didominasi pelanggan yang bertempat tinggal diatas 5 km dari counter penjualan. Karakteristik ini bersesuaian dengan faktor jenis pekerjaan pelanggan, yang pada klaster 1 untuk pelanggan berstatus siswa dan mahasiswa dan klaster 2 untuk pelanggan yang telah bekerja. Pelanggan dengan pekerjaan siswa dan mahasiswa cenderung bertransaksi pada counter yang dekat dengan tempat tinggal, bisa dikarenakan faktor kepemilikan kendaraan ataupun keamanan perjalanan. Pada

klaster 2 yang beranggotakan pelanggan yang telah bekerja, jarak ke counter bukan menjadi pertimbangan dikarenakan faktor usia.

#### 4. Kesimpulan

Hasil yang diharapkan dari penelitian ini adalah diperolehnya rumusan karakteristik kelompok pelanggan sehingga dapat digunakan untuk penyusunan strategi penjualan. Berdasarkan hasil yang diperoleh dari proses segmentasi dan analisis, diketahui bahwa pelanggan dapat dikelompokkan menjadi 2 klaster berdasarkan atribut yang mendominasi yaitu usia dan jenis pekerjaan pelanggan. Klaster 1 beranggotakan pelanggan yang berusia di bawah 24 tahun dan memiliki pekerjaan sebagai siswa dan mahasiswa. Klaster 2 beranggotakan pelanggan yang berusia di atas 24 tahun yang telah memiliki pekerjaan. Berdasarkan kedua klaster tersebut, jumlah transaksi terbanyak terdapat pada klaster 1 dan waktu transaksi yang paling banyak terjadi pada rentang jam 10 pagi sampai jam 3 siang. Penelitian ini berhasil mengkombinasikan KMC dan DBI serta menganalisis karakteristik pelanggan menggunakan teknik statistik deskriptif sederhana. Penelitian ini masih dapat dikembangkan dengan mengkombinasikan algoritma decision tree untuk menganalisis klaster sehingga diperoleh label yang tepat untuk setiap klasternya.

#### Daftar Pustaka

- [1] T. L. Afandi, B. Warsito, and R. Santoso, "Implementasi K-Medoids Dan Model Weighted-Length Recency Frequency Monetary (W-Lrfm) Untuk Segmentasi Pelanggan Dilengkapi Gui R," *J. Gaussian*, vol. 11, no. 3, pp. 429–438, 2023, doi: 10.14710/j.gauss.11.3.429-438.
- [2] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017, doi: 10.1016/j.neucom.2017.06.053.
- [3] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.
- [4] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.
- [5] Trupti M. Kodinariya and Dr. Prashant R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.
- [6] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021, [Online]. Available: <https://ejournal.bsi.ac.id/ejournal/index.php/evolusi/article/view/10428/4839>.
- [7] D. Triyansyah and D. Fitriana, "Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing," *J. Telekomun. dan Komput.*, vol. 8, no. 3, p. 163, 2018, doi: 10.22441/incomtech.v8i3.4174.
- [8] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, *Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature*, vol. 33, no. 11. 2021.
- [9] S. Sajidah, R. Herdiana, and D. Solihudin, "Segmentasi Pelanggan Salon Nuii Beauty Glow Menggunakan K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, pp. 558–566, 2023, doi: 10.36040/jati.v7i1.6333.
- [10] B. Jumadi Dehotman Sitompul, O. Salim Sitompul, and P. Sihombing, "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1235, no. 1, 2019, doi: 10.1088/1742-6596/1235/1/012015.
- [11] D. U. E. Saputri, T. Hidayat, and S. Masturoh, "Pengelompokan Kebutuhan Air Bersih Di Indonesia Periode 2012-2017 Menggunakan Algoritma K-Means Clustering of Clean Water Needs in Indonesia for the 2012-2017 Period Using the K-Means Algorithm," *J. SISFOTENIKA*, vol. 12, no. 2, pp. 203–212, 2022.
- [12] N. Puspitasari, A. A. Maulana, and F. Alameka, "K-Means untuk Klasterisasi Daerah Rawan Penyakit Demam Berdarah K-Means for Clustering of Dengue Fever Prone Areas," *J. SISFOTENIKA*, vol. 13, no. 1, pp. 40–52, 2023.
- [13] Nurindah ; Rachmad Fitriyanto. (2023). customer satisfaction questionnaires 2023[Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/6205835>