Customer Segmentation Menggunakan Algoritma K-Means *Cluster* pada Halal Mart Semarang

Alamsyah[1], Abdul Kholiq[2], Rizka Nur Pratama[3]

[1]Jurusan Ilmu Komputer, FMIPA, Universitas Negeri Semarang
[2]Jurusan Bimbingan Konseling, FIP, Universitas Negeri Semarang
[3]PT Berau Coal, Kalimantan Timur
Email: [1]alamsyah@mail.unnes.ac.id, [2]abdulkholiq@mail.unnes.ac.id, [3]rzkprtma@students.unnes.ac.id

Abstrak

Segmentasi pelanggan bertujuan untuk mengelompokan pelanggan yang memiliki kesamaan karakteristik. Segmentasi pelanggan diperlukan untuk dapat mempertahankan pelanggan lama dengan melakukan perencanaan pelayanan yang paling tepat untuk setiap pelanggan, sehingga dapat menguntungkan perusahaan. *Clustering* merupakan teknik data mining yang dapat membagi kelompok berdasarkan kesamaan karakteristiknya dalam satu *cluster*. Algoritma K-Means *cluster* merupakan salah satu metode *clustering* yang sangat populer dan banyak dipelajari untuk meminimalkan kesalahan *clustering*. Metode Elbow digunakan untuk meningkatkan kinerja algoritma K-Means dengan memperbaiki kelemahan dari algoritma K-Means yaitu membantu untuk memilih nilai k yang optimal untuk digunakan saat *clustering*. Penelitian ini menggunakan data *history* transaksi penjualan di Halal Mart Semarang. Halal Mart merupakan salah satu cabang perusahaan dagang yang berada di Semarang, berfokus menjual produk-produk herbal, produk kecantikan dan produk kebutuhan rumah tangga sehari-hari. Penelitian ini menghasilkan 3 *cluster* dengan nilai *Sum of Square Error* sebesar 544,9. Penelitian ini juga melakukan analisis karakteristik terhadap setiap *cluster* yang dihasilkan.
Kata Kunci: Segmentasi pelanggan, *clustering*, algoritma k-means

Abstract

Customer segmentation aims to group customers who have similar characteristics. Customer segmentation is needed to be able to retain old customers by planning the most appropriate service for each customer, so that it can benefit the company. Clustering is a data mining technique that can divide groups based on similar characteristics in one cluster. The K-Means cluster algorithm is one of the most popular and widely studied clustering methods to minimize clustering errors. The Elbow method is used to improve the performance of the K-Means algorithm by correcting the weakness of the K-Means algorithm, which is helping to choose the optimal k value to be used when clustering. This study uses historical data on sales transactions at Halal Mart Semarang. Halal Mart is a branch of a trading company located in Semarang, focusing on selling herbal products, beauty products and products for daily household needs. This study resulted in 3 clusters with a Sum of Square Error value of 544.9. This study also analyzes the characteristics of each resulting cluster.
***Keyword**: Customer segmentation, clustering, k-means algorithm*

1. PENDAHULUAN

Perkembangan Teknologi Informasi (TI) yang semakin hari semakin menunjukkan peningkatannya tentunya selaras dengan kondisi peradaban manusia, diantaranya terjadi pada sektor bidang bisnis [1]. Dalam bisnis, persaingannya menuntut perusahaan untuk bisa memaksimalkan keterampilan yang ada dengan sebaik-baiknya supaya dapat bersaing dengan perusahaan lain [2]. Perusahaan harus bisa memahami dan

memasukkan karakteristik pelanggan menjadi suatu hal yang penting untuk dipertimbangkan [3].

Perubahan kondisi ekonomi masyarakat berpengaruh pada semakin ketatnya persaingan industri bisnis [4]. Seiring dengan ketatnya persaingan industri bisnis pada perusahaan retail yang semakin meningkat, perusahaan retail diharuskan untuk mengalihkan fokus tidak hanya pada *product oriented* melainkan juga harus melaksanakan strategi yang juga berfokus pada *customer oriented* [5]. Dalam melakukan *customer oriented* diharuskan untuk mengetahui karakteristik dari masing-masing pelanggan [6]. Halal Mart Semarang merupakan salah satu cabang perusahaan dagang yang berada di Semarang. Halal Mart Semarang berfokus menjual produk-produk herbal, produk kecantikan dan produk kebutuhan rumah tangga sehari-hari. Segmentasi pelanggan dibutuhkan guna menggolongkan pelanggan dengan karakteristik yang sama. Dalam satu segmentasi pelanggan merupakan suatu kesatuan kelompok pelanggan yang memiliki karakteristik yang sama [7]. Penerapan segmentasi pelanggan dapat membantu Halal Mart Semarang untuk memberikan pelayanan yang sesuai dengan kebutuhan pelanggan.

Data mining digunakan untuk membantu dalam pengambilan keputusan [8]. *Clustering* adalah teknik data mining dengan melakukan pembagian data pada sebuah himpunan ke dalam kelompok-kelompok dengan kesamaan data di suatu kelompok lebih besar dibanding kesamaan data itu dengan data di kelompok yang lain [7]. Metode *clustering* terpopuler dan banyak dipelajari untuk meminimalisir kesalahan *clustering* adalah K-Means *cluster* [9]. Pendekatan K-Means menggunakan strategi *greedy* (serakah) [10-11] sehingga menghasilkan partisi baru dengan menugaskan tiap-tiap pola ke pusat *cluster* paling dekat dan melakukan perhitungan pusat *cluster* baru [12]. K-Means mengelompokkan suatu data yang ditentukan melalui beberapa *cluster* (k *cluster*). Ide yang dimunculkan adalah memberikan definisi nilai pusat *k* (*k* centroid), satu-satu di setiap klasternya. Dalam menempatkan nilai pusat harus dilakukan dengan pintar karena perbedaan lokasi juga menjadi faktor perbedaan hasil [13].

Untuk membantu proses mengelompokkan tiap-tiap kategori pelanggan serta mengetahui tingkatan loyalitas yang dimiliki adalah dengan menggunakan algoritma K-Means [14]. Metode Elbow digunakan untuk meningkatkan kinerja algoritma K-Means dengan memperbaiki kelemahan dari algoritma K-Means yaitu membantu untuk memilih nilai k yang optimal untuk digunakan saat *clustering* [15]. Metode K-Means lebih optimal kinerjanya dengan menambahkan metode Elbow sebagai metode untuk pemilihan nilai *k cluster* [16]. Penelitian ini berfokus pada penerapan algoritma K-Means *cluster* dengan metode Elbow pada segmentasi pelanggan di Halal Mart Semarang.

2. METODE

2.1. Pengumpulan Data

Proses pengumpulan data dilaksanakan untuk menghimpun dataset yang diperlukan. Penelitian ini menggunakan data yang diperoleh dari data *history* transaksi Halal Mart Semarang. Data merupakan kumpulan data *history* transaksi yang dilakukan pelanggan

Halal Mart Semarang pada bulan Januari sampai Agustus 2022. Data terdiri dari 7 atribut dan 1.174 baris. Atribut dari data *history* transaksi Halal Mart Semarang dijelaskan dalam Tabel 1.

Tabel 1. Atribut dataset

No.	Atribut	Tipe Data
1.	Nomor Invoice	Numerik
2.	ID Pembeli	Nominal
3.	Nama Pembeli	Nominal
4.	Tanggal Penjualan	Numerik
5.	Total VP	Numerik
6.	Item	Numerik
7.	Total	Numerik

2.2. Data Cleaning

Menjadi lebih mudah bagi perusahaan untuk menyimpan dan memperoleh data dalam jumlah besar. Kumpulan data ini dapat memfasilitasi keputusan yang lebih baik membuat, analitik yang lebih kaya, dan semakin banyak, menyediakan data pelatihan untuk data mining [17]. Namun, kualitas data tetap menjadi yang utama, dan data kotor dapat menyebabkan keputusan yang salah dan analisis yang tidak dapat diandalkan. Contoh kesalahan umum termasuk nilai yang hilang, salah ketik, format campuran, entri yang direplikasi dari entitas dunia nyata yang sama, dan pelanggaran aturan bisnis. Analis harus mempertimbangkan efek data kotor sebelum membuat keputusan [18]. Pada penelitian ini, data dilakukan tahapan pembersihan data dari missing values dan atribut-atribut yang tidak diperlukan. Pada tahapan ini dilakukan penghapusan kolom atribut Nomor Invoice, ID Pembeli, Nama Pembeli dan Tanggal Penjualan.

2.3. Standarisasi Data

Dalam proses ini, dilaksanakan standarisasi data pada dataset. Pada penelitian ini, standarisasi data dilakukan dengan menggunakan *Standard Scaler Standardization*. *Standard Scaler* merupakan metode *pre-processing* dimana metode tersebut akan melakukan standarisasi fitur dengan menghapus rata-rata dan menskalakan unit varian. Proses tersebut akan dilakukan pada setiap fitur pada sampel. *Pre-processing* ini dilakukan untuk mencegah adanya data yang memiliki nilai terlalu besar dibanding dengan yang lain yang akan dapat mengakibatkan proses training tidak sesuai dengan keinginan. *Standard Scaler* menghapus mean (terpusat pada 0) dan melakukan skala ke variasi (deviasi standar = 1), dengan asumsi data terdistribusi normal (gauss) pada seluruh fitur [19]. Formula yang digunakan pada standarisasi standar scaler ditunjukkan pada Persamaan (1).

$$\boldsymbol{z} = \frac{(\boldsymbol{x}-\boldsymbol{u})}{\boldsymbol{s}} \tag{1}$$

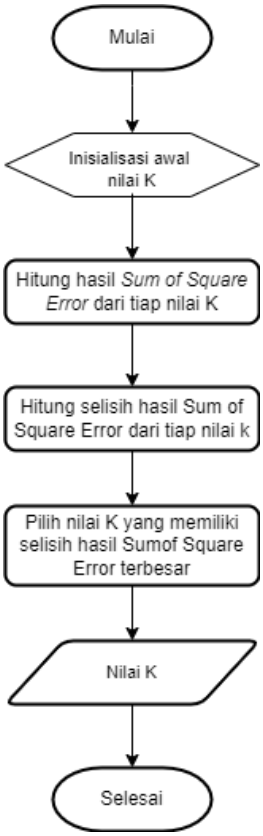
Dengan keterangan,
u = rata-rata nilai sampel
s = standar deviasi dari data training

Pemusatan dan *scaling* terjadi secara independen pada setiap fitur dengan menghitung statistik yang relevan pada sampel dalam dataset hasil *training*. Rata-rata dan standar

deviasi kemudian disimpan untuk digunakan pada data selanjutnya menggunakan transformasi.

2.4. Elbow Method

Metode Elbow digunakan untuk mendapatkan nilai *cluster* terbaik menggunakan cara melakukan pemilihan nilai *cluster* selanjutnya melakukan penambahan nilai *cluster* itu. Metode ini bekerja dengan melihat perbandingan hasil persentase setiap *cluster* yang berbentuk siku di sebuah titik guna menghasilkan informasi penentuan nilai k terbaik [20]. Hasil persentase dari setiap nilai *cluster* dapat diinformasikan sebagai sumber informasi dengan menggunakan grafik. Pada grafik akan memunculkan beberapa nilai k yang paling banyak turunnya serta hasil nilai k akan mengalami penurunan dengan sedikit demi sedikit hingga hasil nilai k ini tidak terjadi penurunan kembali. Metode Elbow memiliki kelemahan yaitu pada penentuan identifikasi titik siku yang tidak selalu dapat dengan mudah diidentifikasi [21]. Flowchart metode Elbow dijelaskan pada Gambar 1.



Gambar 1. Flowchart metode elbow

2.5. Sum of Square Error

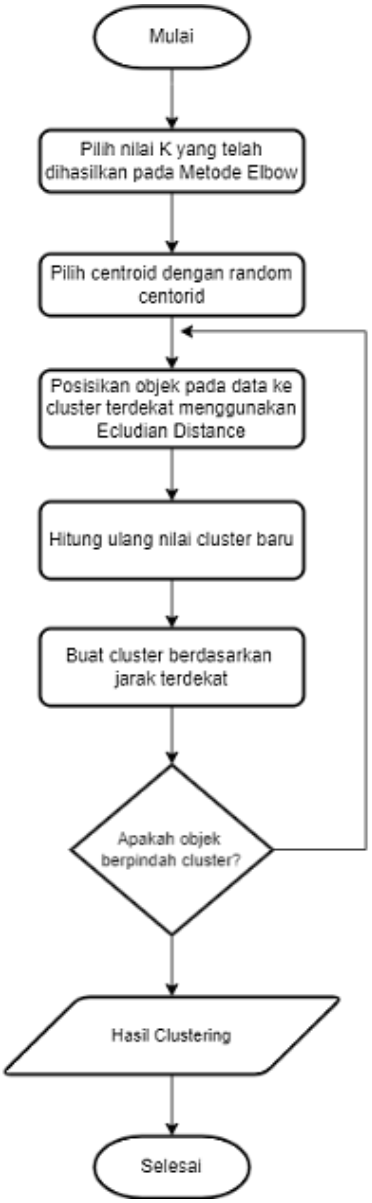
SSE adalah salah satu metode yang digunakan untuk evaluasi dalam mengukur keseragaman antar variabel di dalam satu *cluster*. Metode SSE hanya menggunakan informasi yang terdapat pada objek, sehingga SSE merupakan termasuk ke dalam kriteria kualitas internal. Semakin besar selisih nilai SSE antar variabel maka semakin bagus hasil *clustering*-nya. Rumus yang digunakan dalam mencari nilai SSE disajikan dalam Persamaan (2) [22].

$$SSE = \Sigma_{i=1}^{k} \Sigma_{x \epsilon C_i} dist^2(m_i, x) \quad (2)$$

Keterangan:
k = jumlah *cluster*
C_i = *cluster* ke-i
m_i = *cluster* ke-i
x = data yang ada pada setiap *cluster*

2.6. Algoritma K-Means *Cluster*

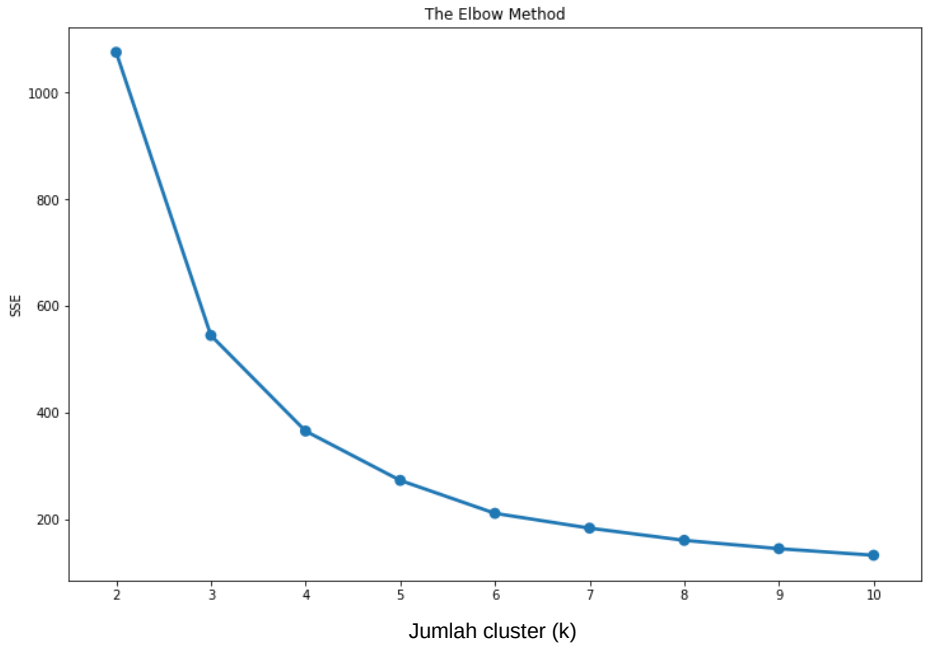
Algoritma K-means adalah algoritma *clustering* untuk mengelompokkan data ke dalam kelompok-kelompok tertentu. Untuk mengambil data pada algoritma ini dengan tidak menggunakan label kelas (*unsupervised learning*). Proses *clustering* K-Means dengan tidak mengetahui target kelasnya membagi kelompok data-data yang menjadi masukannya secara mandiri. Pada tiap-tiap *cluster* memiliki titik pusat (centroid) yang mewakili *cluster* sebagaimana disajikan pada Gambar 2 [23].



Gambar 2. Flowchart algoritma k-means

3. HASIL DAN PEMBAHASAN

Dari penelitian yang telah dilakukan menggunakan algoritma K-Means menghasilkan 3 *cluster*. Pemilihan nilai *k cluster* pada penelitian ini menggunakan metode Elbow untuk memperoleh nilai *k cluster* yang optimal. Hasil dari metode Elbow ditunjukkan pada Gambar 3.



Gambar 3. Grafik chart metode Elbow

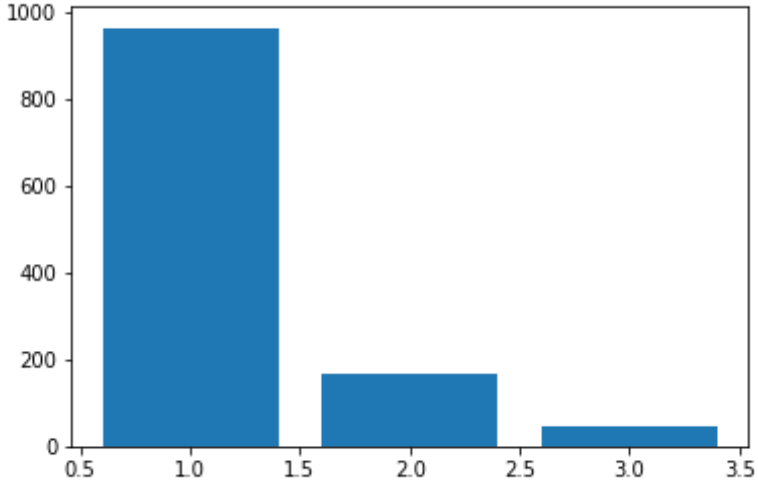
Pemilihan nilai *k cluster* optimal dilihat melalui grafik yang menunjukan titik yang membentuk siku dan memiliki nilai SSE terbesar. Adapun nilai SSE dari setiap nilai k *cluster* ditunjukkan pada Tabel 2.

Tabel 2. Hasil nilai SSE

Jumlah *cluster*	Nilai SSE
2	1.075,4
3	544,9
4	365,7
5	273,3
6	211,6
7	183,8
8	161,0
9	145,3
10	133,0

Berdasarkan Tabel 1 diketahui bahwa nilai k *cluster* yang paling optimal berdasarkan titik yang membentuk siku dan memiliki selisih nilai SSE yang paling besar adalah nilai *k cluster* = 3.

Tahapan *clustering* selanjutnya dilakukan setelah mendapatkan nilai *k cluster*. Hasil dari *clustering* yang telah dilakukan pada penelitian ini ditunjukkan pada Gambar 4.



Gambar 4. Jumlah anggota tiap *cluster*

Setiap *cluster* yang telah dihasilkan dilakukan analisis terhadap variabel Total VP, Item, dan Total. Analisis *cluster* dilakukan untuk mendapatkan informasi terhadap rata-rata, dan Total VP, Item, dan Total yang dilakukan oleh pelanggan Halal Mart, yang ditunjukkan pada Tabel 3. Analisis *cluster* terhadap range dari variabel Total VP, Item, dan Total ditunjukkan pada Tabel 4.

Tabel 3. Average tiap *cluster*

Cluster	Average Total VP	Average Item	Average Total
1	60.296	4	201.333
2	383.430	24	1.268.068
3	874.298	58	2.922.225

Tabel 4. Range tiap *cluster*

Cluster	Range Total VP	Range Item	Range Total
1	0 – 245.000	0 - 30	0 – 780.000
2	167.000 – 650.000	4 – 57	594.000 – 2.108.000
3	454.000 – 1.496.500	30 - 96	1.754.000 – 5.193.000

Hasil jugammlah *cluster* berdasarkan penerapan metode Elbow dan algoritma K-Means menghasilkan 3 *cluster*, dalam proses menentukan jumlah *cluster* bisa membuat pemisahan golongan kelompok *customer* berdasarkan karakteristik dari *history* transaksi yang telah dilakukan. Bentuk karakter atas *cluster* yang dihasilkan membagi menjadi 3 yaitu *cluster* 1 dengan karakteristik pelanggan yang melakukan transaksi paling sedikit memberikan keuntungan, *cluster* 2 merupakan pelanggan dengan karakteristik transaksi menengah, dan *cluster* 3 merupakan karakteristik pelanggan yang paling banyak memberikan keuntungan. Sehingga pelanggan yang berada pada *cluster* 3 dapat dijadikan prioritas bagi Halal Mart Semarang.

4. SIMPULAN

Berdasarkan hasil penelitian dan pembahasan mengenai *clustering* menggunakan algoritma K-Means *cluster* dengan penerapan metode Elbow pada pemilihan nilai k *cluster* pada data *history* transaksi pelanggan Halal Mart Semarang menghasilkan 3 *cluster* dengan nilai SSE sebesar 544,9. Setiap *cluster* menginformasikan karakteristik pelanggan dalam melakukan transaksi pembelian yang berbeda-beda. Bentuk karakter atas *cluster* yang dihasilkan membagi menjadi 3 yaitu *cluster* 1 dengan karakteristik pelanggan yang melakukan transaksi paling sedikit memberikan keuntungan, *cluster* 2 merupakan pelanggan dengan karakteristik transaksi menengah, dan *cluster* 3 merupakan karakteristik pelanggan yang paling banyak memberikan keuntungan. Sehingga pelanggan yang berada pada *cluster* 3 dapat dijadikan prioritas bagi Halal Mart Semarang. Penelitian ini dapat ditarik kesimpulan bahwa penerapan algoritma K-Means *cluster* dapat bekerja secara optimal untuk menghasilkan segmentasi pelanggan pada Halal Mart Semarang.

5. REFERENSI

[1] B. E. Adiana, I. Soesanti, A. E. Permanasari, J. G. No, J. G. No, and J. G. No, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering," *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018.

[2] H. Zhao and C. He, "Objective Cluster Analysis in Value-Based Customer Segmentation Method," in *2009 Second Int. Workshop Knowl. Discov. Data Min.*, 2009, pp. 484–487.

[3] Y. Chen, G. Zhang, D. Hu, and S. Wang, "Customer Segmentation in Customer Relationship Management Based on Data Mining," in *Int. Conf. Program. Lang. Manuf.*, 2006, pp. 288–293.

[4] I. Soesanti, "Web-Based Monitoring System on The Production Process of Yogyakarta Batik Industry," *J. Theor. Appl. Inf. Technol.*, vol. 87, no. 1, pp. 146–152, 2016.

[5] N. W. Wardani *et al.*, "Prediksi Customer Churn dengan Algoritma Decision Tree C4.5 Berdasarkan Segmentasi Pelanggan Pada Perusahaan Retail," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 1, no. 1, pp. 16–24, 2018.

[6] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "Applying Data Mining and RFM Model to Analyze Customers' Values of a Veterinary Hospital," in *2016 Int. Symp. Comput. Consum. Control (IS3C)*, 2016, pp. 481–484.

[7] J. Wu and Z. Lin, "Research on Customer Segmentation Model by Clustering," in *Proc. 7th int. conf. Electron. commer.*, 2005, pp. 316–318.

[8] C.-H. Cheng and Y.-S. Chen, "Classifying The Segmentation of Customer Value Via RFM Model and RS Theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, 2009.

[9] M. A. Berry and G. S. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management*. Emerald Group Publishing Limited, 2000.

[10] G. I. Sampurno, E. Sugiharti, and A. Alamsyah, "Comparison of Dynamic Programming Algorithm and Greedy Algorithm on Integer Knapsack Problem in Freight Transportation," *Sci. J. Informatics*, vol. 5, no. 1, p. 49, 2018.

[11] Alamsyah and I. T. Putri, "Penerapan Algoritma Greedy Pada Mesin Penjual Otomatis (Vending Machine)," *Sci. J. Informatics*, vol. 1, no. 2, pp. 201–209, 2014.

[12] L. Ye, C. Qiu-ru, X. Hai-xu, L. Yi-jun, and Y. Zhi-min, "Telecom Customer Segmentation with K-Means Clustering," in *2012 7th Int. Conf. Comput. Sci. Educ. (ICCSE)*, 2012, pp. 648–651.

[13] N. Kurinjivendhan and K. Thangadurai, "Modified K-Means Algorithm and Genetic Approach for Cluster Optimization," in *2016 Int. Conf. Data Min. Adv. Comput. (SAPIENCE)*, 2016, pp. 53–56.

[14] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[15] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 975–8887, 2014.

[16] O. Dogan, E. Ayçin, and Z. Bulut, "Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry," *Int. J. Contemp. Econ. Adm. Sci.*, vol. 8, no. 1, pp. 1–19, 2018.

[17] T. M. Kodinariya and P. 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.

[18] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The Application of K-means Clustering for Province Clustering in Indonesia of The Risk of The COVID-19 Pandemic Based on COVID-19 Data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022.

[19] Y.-S. Chen, C.-H. Cheng, C.-J. Lai, C.-Y. Hsu, and H.-J. Syu, "Identifying Patients in Target Customer Segments Using A Two-Stage Clustering-Classification Approach: A Hospital-Based Assessment," *Comput. Biol. Med.*, vol. 42, no. 2, pp. 213–221, Feb. 2012.

[20] A. M. Hughes, *Strategic Database Marketing*. McGraw-Hill Pub. Co., 2005.

[21] P.-N. Tan, M. Steinbach, and V. Kumar, *Data Mining Introduction*. Beijing: People's Posts Telecommun. Publ. House, 2006.

[22] S. Asteriadis, K. Karpouzis, N. Shaker, and G. N. Yannakakis, "Towards Detecting Clusters of Players using Visual and Gameplay Behavioral Cues," *Procedia Comput. Sci.*, vol. 15, pp. 140–147, 2012.

[23] T. Caliński and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.