
SEGMENTASI PELANGGAN MENGGUNAKAN METODE K-MEANS CLUSTERING BERDASARKAN MODEL RFM (*RECENCY, FREQUENCY, MONETARY*)

Muhammad H. Anshary^{1*}, Oni Soesanto², Ayatullah³

^{1,2}Program Studi Statistika Fakultas MIPA Universitas Lambung Mangkurat, Kalimantan Selatan, Indonesia

³Dinas Komunikasi dan Informatika Pemerintah Provinsi Kalimantan Selatan, Indonesia

*e-mail corresponding author: hfdzanshary@gmail.com

Abstract

Companies or entrepreneurs must better understanding customers data in all aspects, including detecting similarities and differences among customers, predicting their behavior, and offering better options and opportunities to customers. Customer segmentation is carried out to obtain this information, which is part of CRM (Customer Relationship Management). One of the general models in the application of customer segmentation is the RFM (Recency, Frequency, and Monetary) model. This research method uses a combination of the RFM model and clustering. RFM is used as a description of customer behavior in conducting transactions. Clustering is a process that is widely used and is designed to categorize data. Clustering uses the K-Means Algorithm to determine the number of clusters using the Elbow and Silhouette methods. The application of RFM analysis and the K-Means resulted in two customer segments, namely potential customers and non-potential customers. Potential customers have the characteristics of frequent transactions and also large expenses. Non-potential customers have the characteristics of infrequent transactions and also standard expenses.

Keywords: Customer Segmentation, RFM Model, K-Means Clustering

1. PENDAHULUAN

Dalam skenario kompetitif zaman sekarang, perusahaan mulai mengevaluasi dan mengelola pengalaman pelanggan melalui teknik-teknik pemasaran relasional dalam rangka menjalin dan meningkatkan hubungan dengan pelanggan. Perusahaan atau pelaku bisnis perlu memahami data pelanggan dengan lebih baik dalam semua hal aspek seperti mendeteksi persamaan dan perbedaan di antara para pelanggan, memprediksi perilaku mereka, mengusulkan pilihan dan kesempatan yang lebih baik untuk pelanggan menjadi sangat penting untuk keterlibatan hubungan pelanggan dengan perusahaan. Segmentasi pelanggan dari data mereka menjadi penting dalam konteks hubungan perusahaan usaha dengan pelanggan [5].

Segmentasi pelanggan merupakan bagian dari CRM (*Customer Relationship Management*) yang secara umum membagi informasi menjadi empat bagian yaitu *demographic information, geographical information, psychographic*, dan *behavioral data*. Salah satu model umum dalam penerapan segmentasi pelanggan adalah model RFM (*Recency, Frequency, dan Monetary*). Data dibagi berdasarkan tiga dimensi yaitu *Recency* dengan tujuan mengetahui jeda waktu atau hari pelanggan berdasarkan terakhir waktu atau hari transaksi, kemudian *Frequency* dengan tujuan mengetahui jumlah kedatangan atau kunjungan pelanggan melakukan pembelian dan terakhir adalah *Monetary* dengan tujuan mengetahui jumlah keuntungan perusahaan berdasarkan uang dari pelanggan [4].

Dalam melakukan segmentasi pelanggan, umumnya dapat menggunakan metode *clustering* dengan tujuan dapat memberikan informasi dengan jelas dan kredibel. Teknik *data mining*, khususnya teknik *clustering*, memungkinkan semua pelanggan untuk dibagi menjadi beberapa kelompok (*cluster*) berdasarkan beberapa kesamaan di antara mereka. Algoritma K-Means adalah algoritma klasik untuk menyelesaikan masalah *clustering*. K-Means memiliki metode iteratif yang sederhana, seperti membagi *dataset* tertentu menjadi sejumlah *cluster* tertentu, atau menggunakan metode *Elbow* dan *Silhouette* untuk menentukan jumlah *cluster* yang optimal [11].

Berdasarkan latar belakang tersebut maka terdapat dua tujuan untuk melakukan penelitian ini yaitu mengetahui hasil segmentasi pelanggan menggunakan kombinasi *clustering* dengan algoritma K-Means dan menentukan karakteristik dari hasil segmentasi klasterisasi dari model RFM (*Recency, Frequency* dan *Monetary*) yang terbentuk.

2. TINJAUAN PUSTAKA

2.1 Segmentasi Pelanggan

Proses mengkategorikan pelanggan dengan karakteristik heterogen ke dalam kelompok yang jelas dan nyata, sedangkan karakteristik homogen dimasukkan ke kelompok yang berdasarkan atribut umum merupakan definisi dari segmentasi pelanggan [6]. Segmentasi pelanggan dianggap sebagai metode yang efektif untuk mengembangkan strategi pemasaran yang berbeda berdasarkan karakteristik pelanggan. Segmentasi pelanggan yang efektif berkontribusi dalam meningkatkan tidak hanya kepuasan pelanggan, tetapi juga keuntungan yang diharapkan dari organisasi atau pelaku bisnis [1].

2.2 Customer Value

Analisis dari *customer value* berguna untuk menafsirkan perilaku pelanggan dari luas sumber data yang tidak berarti. *Customer value* diartikan sebagai pendapatan laba bersih per saat ini [8]. Jadi nilai pelanggan didasarkan pada masa lalu dan potensi keuntungan serta probabilitas pemberhentian langganan oleh pelanggan.

2.3 Model RFM

RFM adalah metode umum yang ditujukan untuk mengidentifikasi perilaku pelanggan berdasarkan karakteristik perilaku pelanggan secara *real-time* [10]. Berikut adalah penjelasan dari variabel RFM.

1. *Recency*

Recency adalah perbedaan waktu antara pembelian terakhir pelanggan atau konsumen dengan waktu saat ini. Sehingga didapatkan persamaan 2.1 sebagai berikut ini.

$$Recency = Max(t_i) - t_i, \quad i = 1, 2, 3, \dots, n \quad (1)$$

Dimana t adalah tanggal transaksi pelanggan dan i adalah indeks pelanggan.

2. *Frequency*

Frequency adalah jumlah total pembelian atau kunjungan dari pelanggan atau konsumen selama periode waktu tertentu. Sehingga didapatkan persamaan 2.2 sebagai berikut ini.

$$Frequency = \sum k_t \quad (2)$$

Dimana k_t adalah kedatangan pelanggan pada waktu tertentu.

3. *Monetary*

Monetary adalah uang yang dikeluarkan oleh pelanggan atau konsumen selama periode waktu tertentu.

$$Monetary = \sum_{p=1}^p H_p \times m_{pi}, i = 1, 2, 3, \dots, n \quad (3)$$

Dimana H adalah harga item produk, p adalah indeks pelanggan dan m adalah kuantitas item yang dibeli oleh pelanggan.

2.4 *Standardization Scale*

Atribut RFM memiliki satuan yang berbeda, sehingga dilakukan standarisasi data. Sesuai dengan tujuan standarisasi yaitu untuk menyesuaikan ukuran (*magnitude*) dan bobot relatif dari variabel input. Perhitungan standarisasi menggunakan persamaan skor-z [9].

$$x' = \frac{x - x_{mean}}{\sigma} \quad (4)$$

Dimana;

- x' : hasil nilai standarisasi
- x : nilai yang akan ditransformasi dalam atribut
- x_{mean} : rata-rata nilai atribut yang akan ditransformasi
- σ : standar deviasi atribut yang akan ditransformasi

2.5 *K-Means Clustering*

Algoritma K-Means adalah algoritma klasik untuk memecahkan masalah *clustering*. Pada K-Means terdapat metode iterasi sederhana untuk mempartisi dataset yang diberikan ke dalam sejumlah *cluster* yang ditentukan oleh pengguna. Prosedur algoritma *K-Means* sebagai berikut [3].

1. Tentukan banyaknya *cluster*.

Untuk menentukan jumlah *cluster* atau K dapat dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin dicetuskan menjadi penentuan berapa jumlah *cluster* yang akan digunakan.

2. Tentukan titik *centroid* k (pusat *cluster*) secara acak.
3. Hitung jarak setiap titik ke pusat *cluster*, jarak antara satu data dengan satu *cluster* lainnya akan menentukan data mana yang masuk ke *cluster* yang mana.

Perhitungan jarak antara data dan pusat *cluster* menggunakan *Euclidian Distance* dengan rumus:

$$D(i, j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \dots + (X_{ki} - X_{kj})^2} \quad (5)$$

Dimana;

$D(i, j)$: jarak dari data ke- i ke pusat *cluster* j

X_{ki} : koordinat data

X_{kj} : koordinat *centroid*

4. Kelompokkan data berdasarkan kedekatan dengan *centroid* kemudian perbaharui nilai *centroid* baru dengan lokasi dari pusat *cluster* menggunakan persamaan:

$$\mu_j(t + 1) = \frac{1}{N_{sj}} \sum_{j \in s_j} X_j \quad (6)$$

Dimana;

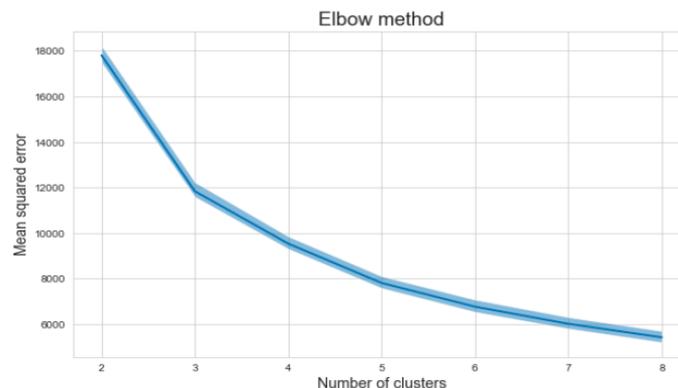
$\mu_j(t + 1)$: *centroid* baru pada iterasi ke- $(t+1)$

N_{sj} : banyaknya data pada *cluster* s_j

5. Lakukan langkah 2 sampai 4 sampai anggota tiap *cluster* tidak ada yang berubah.

2.6 Elbow Method

Metode *elbow* merupakan salah satu cara untuk menentukan nilai K atau jumlah *cluster* yang optimal [2]. Metode *elbow* menggunakan *sum of squared error* (SSE) untuk memilih nilai K atau jumlah *cluster* yang ideal berdasarkan jarak antara titik data dan *cluster* yang ditetapkan. Nilai K atau jumlah *cluster* akan dipilih ketika SSE mulai mendatar dan terlihat titik belok. Ketika divisualisasikan grafik ini akan terlihat seperti siku, seperti nama metodenya.



Gambar 1. Grafik *elbow method*

2.7 Silhouette Index

Silhouette merupakan salah satu dari berbagai metode untuk mengevaluasi hasil *clustering*. Sehingga hasil dari perhitungan dari *silhouette index* lebih dapat diterima dan digunakan untuk mengevaluasi hasil *clustering*. Titik data akan sangat mirip dengan titik data lainnya pada *cluster* yang sama jika nilai koefisien jika nilainya mendekati angka 1. Namun sebaliknya, jika nilainya mendekati -1 maka titik data tersebut tidak mirip dengan titik data di-*cluster*-nya. Perhitungan koefisien *silhouette* dapat dihitung dengan persamaan berikut [7].

$$S_{x(i)} = \frac{b_{x(i)} - a_{x(i)}}{\max(a_{x(i)}, b_{x(i)})} \quad (7)$$

Dimana;

$x(i)$: data yang berada pada *cluster*, $i = 1, 2, 3, \dots, n$.

$a_{x(i)}$: rata-rata jarak x_i dengan semua data yang berada pada *cluster* yang sama.

$b_{x(i)}$: jarak terdekat x_i dengan semua data yang berada pada *cluster* yang tidak sama.

3. METODE PENELITIAN

Penelitian ini akan dilaksanakan melalui tahapan sebagai berikut:

- a. Melakukan *preprocessing* data Pada tahap ini melakukan *data cleaning* dan perubahan tipe data.
- b. Statistika Deskriptif. Analisis deskriptif data berguna untuk memberikan gambaran mengenai data dan menampilkan penyajian data penelitian berguna untuk menyampaikan gambaran mengenai data.
- c. Pembentukan Model RFM. Data dihitung menyesuaikan model RFM yaitu *Recency Frequency, dan Monetary*
- d. Standarisasi Hasil RFM. Standarisasi setiap variabel input secara terpisah dengan mengurangi *mean* dan membaginya dengan standar deviasi sehingga menghasilkan distribusi agar memiliki *mean* nol dan standar deviasi satu.
- e. *Clustering* dengan K-Means. Proses clustering menggunakan algoritma K-Means dengan nilai K atau jumlah *cluster* acak berdasarkan data input dari hasil RFM.
- f. Penentuan dan Verifikasi Optimal *Cluster*. Menentukan dan memverifikasi nilai *cluster k* yang optimal menggunakan metode *Elbow* dan *Silhouette*.
- g. Analisis Karakteristik Pelanggan. Pada tahap ini akan dilakukan analisa karakteristik pelanggan berdasarkan hasil segmentasi.
- h. Interpretasi hasil dan kesimpulan.

4. HASIL DAN PEMBAHASAN

4.1 *Preprocessing Data*

Penelitian menggunakan data sekunder yang didapatkan dari situs *UCI Machine Learning Repository* Data yang digunakan adalah *Online Retail* yang merupakan kumpulan data yang berisi 541.909 transaksi yang terjadi antara 1 Desember 2010 sampai dengan 9 Desember 2011. Data tersebut memiliki 541909 data dengan 8 atribut data yaitu, *InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, dan Country*.

Sebelum data dipakai untuk analisis, data tersebut terlebih dahulu akan dilakukan proses *preprocessing*. Tujuannya adalah untuk menghasilkan data dalam keadaan format yang konsisten dan dapat menghasilkan informasi yang valid. Pada kasus ini ketentuan data yang akan dihapus adalah yang memiliki *null/missing value* dan memiliki keterangan transaksi yang dibatalkan disajikan pada Tabel 1 berikut.

Tabel 1. *missing* data yang dihapus

Invoice No	Invoice Date	Stock Code	Description	Quantity	Unit Price	Customer ID	Country
536414C	2010-12-01 11:52:00	22139	NaN	56	0	NaN	United Kingdom
536544	2010-12-01 14:32:00	21773	DECORATIVE ROSE BATHROOM BOTTLE	1	2.51	NaN	United Kingdom
...
58162	2011-12-09 11:32:00	37811	S/4 CACTUS CANDLES	1	2.15	NaN	United Kingdom
58162	2011-12-09 11:32:00	37812	S/6 CACTUS CANDLES	1	2.15	NaN	United Kingdom

4.2 Statistika Deskriptif

Hasil analisis statistika deskriptif akan dibagi berdasarkan jenis data kualitatif dan kuantitatif, berikut pada Tabel 2 analisis statistika deskriptif yang didapatkan.

Tabel 2. Statistika deskriptif

Data Kualitatif	Invoice No	Stock Code	Description	Customer ID	Country
Jumlah data	392122	392122	392122	392122	392122
Jumlah data <i>unique</i>	17916	3661	3868	4226	37
Modus data	576339	85123A	WHITE HANGING HEART T- HOLDER	17841	United Kingdom
Frekuensi	540	1982	1975	7843	350074

Data Kuantitatif	Quantity (pcs)	Unit Price (£)
Rata-rata	10.18	2.89
Standar deviasi	28.22	3.67
Minimal	1	0.001
Maksimal	12540	165
Jumlah data	392122	392122
Jumlah data <i>unique</i>	155	310

4.3 Pembentukan Model RFM

Selanjutnya data diolah dan dibentuk sesuai atribut R, F dan M yang dihitung menggunakan persamaan 1, 2, dan 3. Hasil pembentukan model RFM dapat dilihat pada Tabel 3 berikut.

Tabel 3. Data dan Hasil Standarisasi RFM

Customer ID	Riil Data RFM			Standarisasi RFM		
	Recency	Frequency	Monetary	Recency	Frequency	Monetary
12346	325	1	0	2.302165	-0.750830	-0.722541
12347	1	181	4060.4	-0.906150	1.057111	1.735232
...
17850	373	1	710.5	2.817377	0.544861	-0.59918
...
18282	7	12	178.05	-0.58628	-0.64034	-0.54967
18287	42	70	1837.28	1.025146	0.936582	-0.89625

4.4 Standarisasi Nilai RFM

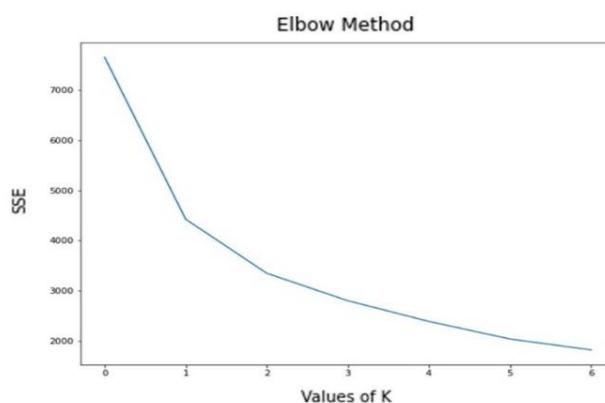
Kemudian setelah itu pada data RFM dilakukan standarisasi dengan menggunakan rumus skor-z atau persamaan 4. Karena masing-masing dari nilai R, F, M memiliki bobot nilai yang berbeda satu sama lain, maka dilakukan standarisasi agar memiliki bobot nilai yang sama secara keseluruhan. Seperti contoh pada nilai M merupakan jumlah uang yang dikeluarkan pelanggan untuk perusahaan dengan satuannya yaitu *Poundsterling* (£). Sehingga selisih yang sangat jauh seperti itulah diperlukan standarisasi agar memudahkan proses analisis data. Hasil perhitungan dapat dilihat Tabel 3.

4.5 Clustering dengan K-Means

Setelah tahapan standarisasi dari nilai RFM telah selesai dilakukan, data tersebut kemudian dilakukan *clustering* dengan algoritma K-Means dengan metode validasi *Elbow Method* dan *Silhouette Coefficient* dalam menentukan jumlah *cluster* atau K yang optimal dengan menguji pembentukan *cluster* yang terbentuk dari *cluster* 2 sampai dengan *cluster* 6. Penentuan jumlah *cluster* atau K yang optimal dari *Elbow Method* adalah titik data yang membentuk siku pada perhitungan SSE, sedangkan dari *Silhouette Index* adalah titik data akan sangat mirip dengan titik data lainnya pada *cluster* yang sama jika nilai koefisien jika nilainya mendekati angka 1. Namun sebaliknya, jika nilainya mendekati -1 maka titik data tersebut tidak mirip dengan titik data di-*cluster*-nya. Hasil kedua metode ini ditunjukkan pada Tabel 4 dan Gambar 2 berikut.

Tabel 4. Skor *Silhouette*

Cluster	Silhouette Score
2	0.5421938592673445
3	0.5098000480555145
4	0.4798250864030441
5	0.4681759649201254
6	0.4173730207610261



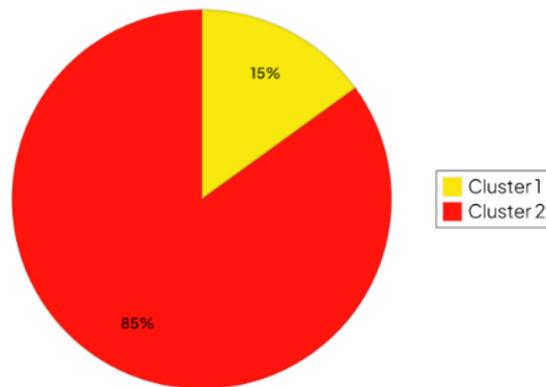
Gambar 2. Hasil *elbow method*

Selanjutnya didapatkan jumlah *cluster* atau K yang optimal dari metode *Elbow* dan *Silhouette* adalah $K = 2$. Sehingga dilanjutkan proses selanjutnya yaitu *clustering* dengan K-Means yang menghasilkan dua segmen pelanggan berdasarkan nilai RFM masing-masing pelanggannya yang dapat dilihat pada Tabel 5.

Tabel 5. *Clustering RFM*

Customer ID	Recency	Frequency	Monetary	Cluster
12346	325	1	0	2
12347	1	181	4060.4	1
...
17850	373	1	710.5	2
...
18282	7	12	178.05	2
18287	42	70	1837.28	2

Pada Tabel 5 ditampilkan *clustering* untuk data RFM dengan K optimal adalah K = 2 untuk masing-masing pelanggan. Didapatkan persentase pelanggan untuk masing-masing *cluster* disajikan pada Gambar 3 berikut.



Gambar 3. Persentase Jumlah anggota tiap *cluster*

Persentase pelanggan pada *cluster* 1 sebesar 15% yang sama dengan 635 pelanggan. Dan jumlah pelanggan pada *cluster* 2 sebesar 3.591 pelanggan atau sekitar 85% dari total pelanggan.

4.6 Analisis Karakteristik Pelanggan

Hasil *clustering* menghasilkan dua segmen pelanggan yaitu *Potential Customers* dan *Non-Potential Customers* yang didasari dari nilai RFM yang sudah terbentuk. Rincian karakteristik kedua segmen pelanggan ini dapat dilihat pada Tabel 6.

Tabel 6. Rincian karakteristik pelanggan

Cluster	1	2
Jumlah anggota	635	3591
Karakteristik pelanggan	<i>Potential Customers</i>	<i>Non-Potential Customers</i>
Recency	0 – 371 hari	0 – 373 hari
Frequency	50 – 710 kali	1 – 260 kali
Monetary	£907.1 – £12206.59	£0 – £4272.76

Berdasarkan Tabel 6. rata-rata *recency* sebesar 23 hari, menunjukkan bahwa pelanggan dengan karakteristik *potential customers* melakukan transaksi dengan jeda waktu yang cukup singkat yaitu kurang dari 1 bulan pada waktu terakhir transaksi

sebelumnya. Dan diketahui juga untuk rata-rata *frequency* adalah 256.17 yang menunjukkan bahwa pelanggan dapat datang atau berbelanja sebanyak 256 kali. Pada *monetary*, rata-rata pengeluaran adalah sebesar £4149.01, menunjukkan bahwa pelanggan dalam satu kali transaksi dapat mengeluarkan uang rata-rata sebesar £4149.01. Hal ini dapat terjadi karena pelanggan dengan karakteristik *potential customers* berbelanja dengan waktu yang cukup sering dan royal dalam pengeluaran.

Rata-rata nilai *recency* sebesar 105 hari, menunjukkan bahwa pelanggan dengan karakteristik *non-potential customers* melakukan transaksi dengan jeda kurang lebih dari 3 bulan pada waktu terakhir transaksi sebelumnya. Sedangkan pada rata-rata nilai *frequency* 44.99 yang menunjukkan bahwa pelanggan dapat datang atau berbelanja sebanyak 45 kali. Pada *monetary* rata-rata pengeluaran adalah sebesar £700.35, menunjukkan bahwa pelanggan dalam satu kali transaksi mengeluarkan rata-rata sebesar £700.35. Hal ini dapat terjadi karena pelanggan dengan karakteristik *non-potential customers* berbelanja dengan waktu yang cukup lama dan membeli barang sesuai dengan kebutuhan atau keperluan mereka.

5. KESIMPULAN

Berdasarkan hasil penelitian pada pembahasan dapat disimpulkan bahwa penelitian ini menggunakan metode *clustering* dengan *K-Means*. Berdasarkan metode *Elbow* dan nilai koefisien *Silhouette* didapatkan nilai $K = 2$ sebagai nilai K yang paling baik untuk melakukan *clustering*, sehingga menghasilkan 2 segmen pelanggan dengan karakteristik yang berbeda. Segmen pelanggan yang terbentuk merupakan lanjutan hasil model RFM dengan acuan rentang waktu terakhir transaksi, frekuensi transaksi, dan jumlah uang yang telah dikeluarkan untuk transaksi. Dari 4226 pelanggan yang dilakukan segmentasi, terdapat 3591 pelanggan yang termasuk dalam karakteristik *Potential Customers* dan terdapat 635 pelanggan yang termasuk pada karakteristik *Non-Potential Customers*. *Potential Customers* mempunyai karakteristik pelanggan yang sering melakukan pembelian dengan pengeluaran yang banyak dalam bertransaksi.

DAFTAR PUSTAKA

- [1] Chen, D., Sain, S. L., & Guo, K. 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing and Customer Strategy Management*, 19(3), 197–208.
- [2] Cui, M. 2020. *Introduction to the K-Means Clustering Algorithm Based on the Elbow Method*.
- [3] Fithri, F. A., & Wardhana, S. 2019. *Cluster Analysis of Sales Transaction Data Using K-Means Clustering at Toko Usaha Mandiri*
- [4] Goyat, S. 2011. The basis of market segmentation: a critical review of literature. In *European Journal of Business and Management* www.iiste.org ISSN (Vol. 3, Issue 9).
- [5] Granata, G. 2020. The Digital Evolution of Consumer Purchasing Methods and the Impact on Retail. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 10(4).

- [6] Hong, T., & Kim, E. 2012. Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, 39(2), 2127–2131.
- [7] Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K., & Kerdprasop, N. (2018). *The silhouette width criterion for clustering and association mining to select image features*. *International Journal of Machine Learning and Computing*, 8(1), 69–73.
- [8] Kotler, P. 2017. *Customer Value Management*. *Journal of Creating Value*, 3(2), 170–172.
- [9] Miuigan, G. W., Cooper, M. C., Milligan, G. W., Milligan, G. W., & Cooper, M. C. 1988. A Study of Standardization of Variables in Cluster Analysis. In *Journal of Classification* (Vol. 5).
- [10] Sohrabi, B., & Khanlari, A. 2007. Customer Lifetime Value (CLV) Measurement Based on RFM Model. In *Iranian Accounting & Auditing Review* (Vol. 14, Issue 47).
- [11] Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37.