 ISSN:2302-8149
e-ISSN:2540-9719

Penerapan Algoritma K-Means dan K-Medoids untuk Segmentasi Pelanggan pada Data Transaksi E-Commerce

The Implementation of K-Means and K-Medoids Algorithm for Customer Segmentation on E-commerce Data Transactions

[1]Romadansyah Siagian*, [2]Pahala Sirait, [3]Arwin Halim
Pusat Komputer, Universitas Medan Area, Medan
Jl. Kolam No 1, Medan Estate, 20223, Indonesia
*e-mail: romadansyahsiagian@gmail.com

(***received*:** 15 Maret 2021, ***revised*:** 22 Juni 2021, ***accepted*:** 16 Februari 2022)

Abstrak

Data transaksi *e-commerce* yang semakin banyak dapat dimanfaatkan perusahaan untuk memberikan informasi yang baru. Data transaksi tersebut dapat mengungkap tentang segmentasi atau kelompok pelanggan berdasarkan kesamaan karakteristik dan perilaku pelanggan masing-masing. Salah satu teknik yang dapat digunakan untuk mensegmentasi pelanggan adalah *Data Mining* dengan menggunakan metode *clustering*. Tujuan penelitian ini adalah menerapkan metode *clustering* pada data transaksi *e-commerce* menggunakan algoritma K-Means dan K-Medoids. Hasil penelitian menunjukkan bahwa algoritma K-Means dan K-Medoids sama-sama menunjukkan hasil *cluster* optimal *k = 3*. Hasil tersebut juga sesuai dengan hasil metode *elbow* dan uji validitas *Davies Bouldin Index* yang menunjukkan bahwa jumlah *cluster* optimal adalah 3 (tiga). Hasil pengujian menunjukkan K-Medoids memiliki performa terbaik dengan nilai ration sebesar 0,337575 dibandingkan K-Means 0,3380724, sehingga K-Medoids digunakan dalam *clustering* data.sebagai *cluster* optimal. Hasil segmentasi pelanggan sesuai *Customer Loyalty Matrix* terdiri dari *core customer, new customers, dan lost customer.*

Kata kunci: Segmentasi Pelanggan, *Clustering*, K-Means, Model LRFM.

Abstract

Nowadays, e-commerce data transactions are commonly used by companies to provide new information. The data transaction can reveal customer segmentation or groups based on the similar characteristics and behavior of each customer. Data Mining is one of technique to conduct the customer segmentation through clustering method. The study aims to applied the clustering method on e-commerce data transactions by using both K-Means and K-Medoids algorithm. The result shows that both algorithms reveal optimum of cluster result with value of k = 3. The results are also indicating the conformity with the elbow method's results and the Davies Bouldin Index validity test which shows that the optimal number of clusters is 3. The test results show that K-Medoids has the best performance with a ration value of 0.337575 compared to K-Means 0.3380724. Hence, K-Medoids are used in data clustering as the optimal cluster. The results of customer segmentation according to the Customer Loyalty Matrix consist of core customers, new customers, and lost customers.

Keywords: *Customer Segmentation, Clustering, K-Means, LRFM Model.*

1 Pendahuluan

Pada tahun 2019 diperkirakan 1,92 miliar orang melakukan transaksi jual beli secara *online* melalui platform *e-commerce* di seluruh dunia dan diperkirakan akan terus meningkat di masa depan [1]. Peningkatan transaksi tersebut memberikan dampak pada semakin banyaknya data transaksi *e-commerce* yang tersimpan. Aset utama perusahaan untuk dipertahankan ialah pelanggan [2]. Perusahaan harus lebih memahami karakteristik, perilaku dan kebutuhan pelanggan yang berbeda-beda.

Caranya dengan menggali informasi dari riwayat data transaski pelanggan *e-commerce* tersebut. Salah satunya dengan mensegmentasi pelanggan dengan tujuan memprediksi dan menargetkan pelanggan potensial guna menarik pelanggan baru, menerapkan strategi pemasaran yang tepat guna, mengevaluasi nilai umur pelanggan, mengenali hubungan antara pelanggan dan perusahaan, serta meningkatkan profitabilitas yang diharapkan perusahaan [3, 4, 5]. Segmentasi tersebut menggunakan metode *clustering* dengan algoritma K-Means dan K-Medoids.

Data transaksi *e-commerce* terlebih dahulu ditranformasikan menjadi model LRFM (*Length*, *Recency*, *Frequency* dan *Monetary*) yaitu model yang memberikan pandangan yang luas dan lebih akurat tentang perilaku pelanggan yang sebenarnya [2, 6], sehingga pengambilan keputusan dan skema strategi promosi menjadi efektif dan efisien [7].

Salah satu metode pada *data mining* adalah *clustering* satu teknik penyelesaian terkait segmentasi [2, 6]. Algoritma K-Means merupakan salah satu algoritma teknik *clustering* yang dapat digunakan analisis *cluster* memiliki kelebihan seperti kecepatan komputasi yang lebih tinggi, mudah diimplementasikan dan dijalankan, bersifat dinamis pada data yang tersebar dan hasil yang diperoleh lebih akurat. Selain itu algoritma K-Medoids juga termasuk kelompok metode *partitional clustering* yang merupakan varian dari metode K-Means. K-Medoids menjadi penyempurnaan metode sebelumnya. Inisialisasi jumlah *cluster k* secara *random* tidak selalu memberikan hasil yang baik dan akurat [8]. Guna mengetahui jumlah *cluster k* yang optimal digunakan metode *Elbow* [5]. Metode uji validitas untuk mengevaluasi hasil penentuan jumlah *cluster k* terbaik menggunakan DBI [9].

Penelitian ini dilakukan untuk menerapan algoritma K-Means dan K-Medoids pada data transaksi *e-commerce* untuk mengetahui segmentasi / kelompok pelanggan perusahaan berdasarkan riwayat transaksi *e-commerce* berdasarkan model LRFM. Tujuan penelitian ini adalah menerapkan metode K-Means dan K-Medoids untuk data transaksi *e-commerce* dan mengetahui metode *cluster* mana yang terbaik dalam *clustering*, sehingga menghasilkan informasi segmentasi pelanggan yang lebih baik bagi perusahaan.

2 Tinjauan Literatur

Penelitian sebelumnya yang berkaitan dengan segmentasi pelanggan dengan menggunakan metode *clustering* dengan algoritma K-Means berdasarkan model LRFM maupun RFM seperti penelitian Marisa, dkk., [2]. Menggunakan Teknik *clustering* dalam mengetahui segmentasi pelanggan dan tahapan penelitian menggunakan algoritma K-Means dengan penentuan *cluster k* terbaik dengan metode *Elbow*. Hasil *cluster* K-Means dengan metode *Elbow* diperoleh nilai *Sum Square Error* (SSE) pada nilai $k = 2$ sebagai *cluster* terbaik. Monalisa 2018 [10]. Melakukan segmentasi pelanggan dengan algoritma K-Means dan metode validasi *Dunn Index* dan *Silhoutte Coefficient* sebagai penentuan jumlah *cluster* optimal dengan hasil nilai *cluster k* terbaik dari kedua metode pada $k = 3$.

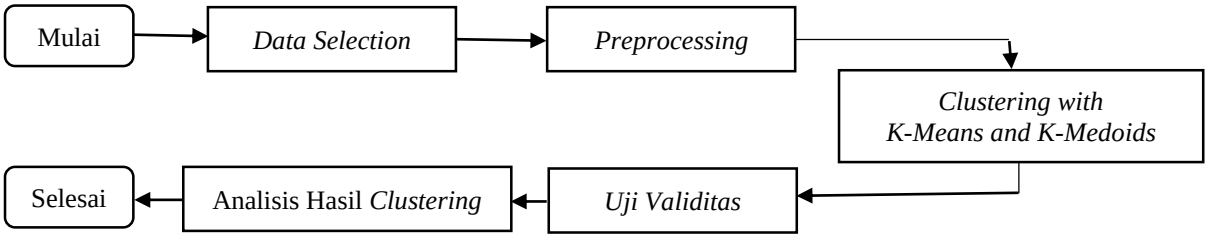
Pada penelitian yang dilakukan Gustriansyah, dkk., [11]. Melakukan segmentasi pasar dengan algortima K-Means dengan evaluasi *cluster* optimal delapan indeks validitas yaitu *Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index,* dan *Krzanowski Index-Lai.* Diperoleh jumlah *cluster* optimal sebanyak $k = 3$ dari validitas indeks *Ratkowski Index, Hubert Index,* dan *Ball-Hall Index cluster dan* dipilih sebagai acuan analisis RPM. Aryuni, dkk [12] menerapkan algoritma K-Means dan K-Medoids untuk segmentasi *customer*, diperoleh nilai k yang paling optimal pada kedua algoritma adalah $k = 3$. Maka dapat disimpulkan dalam pengolahan data, algoritma K-Means memiliki performa yang lebih baik daripada K-Medoids baik dari sisi nilai *average within centroid distance* dan kompleksitas waktu.

Referensi penelitian sebelumnya terkait algoritma K-Means dan K-Medoids memberikan hasil yang cukup beragam, sehingga tidak selalu memberikan hasil baik dan efektif sesuai kondisi permasalahan dan datanya. Sehingga penelitian ini, data transaksi *ecommerce* akan dilakukan *clustering* dengan algortima K-Means dan K-Medoids dan dikombinasikan dengan model LRFM untuk memberikan informasi dan pengetahuan terkait segmentasi pelanggan sehingga perusahaan *e-commerce* mampu lebih mengenali karakteristik dan kelompok pelanggannya guna penyusunan strategi pemasaran.

3 Metode Penelitian

Metode yang digunakan pada penerapan metode *clustering* untuk segmentasi pelanggan terdiri dari tahapan-tahapan yang digambarkan dalam bentuk *flowchart* yang terdapat pada Gambar 1.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1. *Flowchart* Tahapan Penelitian.

3.1 *Data Selection*

Dataset yang digunakan dari transaksi pelanggan *Online Retail* bersumber dari UCI *Machine Learning Repository* pada website https://www.archive.ics.uci.edu/ml/datasets/Online+Retail. *Dataset* tersebut terdiri dari 541909 *instance* dengan 8 atribut dari rentang waktu 01/12/2010 sampai 09/12/2011 [13]. Pemilihan atribut berdasarkan model LRFM sehingga 4 atribut dipilih yang bersesuaian yaitu atribut *customerid*, *quantity*, *invoicedate*, dan *unitprice* sebagaimana terlihat pada Tabel 1.

Tabel 1. Pemilihan Atribut

Atribut	Ya/Tidak	Atribut	Ya/Tidak
InvoiceNo	Tidak	InvoiceDate	Ya
StockCode	Tidak	UnitPrice	Ya
Description	Tidak	CustomerID	Ya
Quantity	Ya	Country	Tidak

3.2 *Preprocessing*

Tahapan dalam *preprocessing* ini terdiri dari beberapa tahapan diantaranya.

a. Data *cleaning* membersihan data tidak valid, data kosong, data *duplicate*, bernilai nilai negatif.
b. Data *transformation* dilakukan pada atribut *customerid*, *quantity*, *invoicedate*, dan *unitprice* agar lebih terukur dan sesuai pada model LRFM sehingga dapat digunakan sebagai atribut *clustering*. Kemudian dilakukan *attribute construction* sesuai atribut LRFM (*length*, *recency*, *frequency*, dan *monetary*) diantaranya [2, 6].
    1) Atribut *length* ialah interval waktu (jumlah hari) antara waktu pembelian awal dan waktu terakhir pembelian pelanggan. Atribut *length* dari *InvoiceDate.*
    2) Atribut *recency* ialah rentang waktu terakhir pelanggan melakukan transaksi pada akhir skala waktu penelitian. Atribut *recency* dari *InvoiceDate.*
    3) Atribut *frequency* ialah berapa kali transaksi pembelian dilakukan oleh pelanggan dalam skala waktu penelitian. Atribut *frequency* dari *InvoiceDate.*
    4) Atribut *monetary* ialah jumlah nominal transaksi untuk setiap pelanggan dalam skala waktu penelitian. Atribut *monetary* dari akumulasi *Quantity* dikali *UnitPrice.*

Hasil *transformasi* menjadi data *input* untuk proses *clustering*, terlebih dahulu di normalisasi tanpa pemberian bobot nilai. Metode Normalisasi Min-Max dengan range yang digunakan pada penelitian ini yaitu nilai antara 0 – 1. Rumus normalisasi min-max adalah sebagai berikut [14].

$$x' = \frac{x - nilai_{min}}{nilai_{max} - nilai_{min}} \quad (1)$$

Kemudian dilakukan pengecekan *outlier* menggunakan rumus rentang interkuartil (IQR) yaitu perbedaan antara persentil ke-75 (Q3) dan persentil ke-25 (Q1) pada kumpulan data menggunakan aturan 1,5 x IQR yaitu nilai di bawah rentang Q1 - (1,5 x IQR) atau di atas rentang Q3 + (1,5 x IQR) adalah pencilan [15, 16]. Rumus interkuartil (IQR) sebagai berikut.

IQR = Q3 – Q1 (2)

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 *Clustering with K-Means and K-Medoids*

Sebelum ke tahapan *clustering* terlebih dahulu menentukan nilai *k* yang optimal. Hal ini dilakukan karena penentuan nilai *k* secara random terkadang kurang tepat dan berpengaruh pada hasil *cluster*, maka dalam penelitian ini *cluster k* ditentukan terlebih dahulu menggunakan metode. Metode *Elbow* memberikan informasi visualisasi perbandingan antara jumlah *cluster* yang membentuk sudut siku pada satu titik grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik dan dengan membandingkan hitungan nilai *Sum Square Error* (SSE) dengan persamaan 3 sebagai berikut [2].

$$SSE = \sum_{K=1}^{K} \sum |x_i - c_k|^2 \quad (3)$$

a. *Clustering with K-Means*

Algoritma K-Means membagi data sejumlah *k cluster* yang sudah ditetapkan diawal secara *random.* Metode K-Means sangat sederhana dimulai dengan pemilihan jumlah *cluster* sebanyak *k* buah. Secara *random k* diambil dari *dataset* sebagai *centroid* yang mewakili suatu *cluster.* Pusat atau titik tengah suatu *cluster* dinamakan *centroid*. Semua data dihitung jaraknya terhadap *centroid* dan setiap data akan menjadi anggota dari sebuah *cluster* yang diwakili oleh *centroid* yang memiliki jarak terdekat dengan data tersebut. Tahap akhir menghitung ulang nilai *centroid* yang diperoleh dari nilai rata-rata dari setiap *cluster* yang ada. Proses pemilihan keanggotaan *cluster* dan perhitungan ulang *centroid* dilakukan terus menerus dan berhenti jika keanggotaan *cluster* tidak mengalami perubahan atau jumlah perulangan yang dilakukan telah melampaui suatu nilai batas tertentu [17]. Prosedur algoritma K-Means sebagai berikut [10, 18].

1. Tentukan jumlah *k*, *k* adalah jumlah *cluster*.
2. Tentukan nilai awal titik pusat *cluster* untuk dilakukannya proses *clustering*.
3. Hitung *Distance Measure* (jarak data) terhadap masing-masing *centroid* menggunakan *Euclidean Distance*.
4. Alokasikan seluruh objek data yang telah dihitung ke dalam masing-masing *cluster*.
5. Tentukan *centroid* baru dengan menggunakan persamaan berikut:

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{n} x_{kj} \quad (4)$$

6. Ulang kembali langkah 3, 4 dan 5 hingga tidak ada lagi anggota cluster yang berpindah ke *cluster* lain.

Perhitungan jarak menggunakan *euclidean distance* merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam *euclidean space* (meliputi bidang *euclidean* dua dimensi, tiga dimensi, atau bahkan lebih) guna mengukur tingkat derajat kemiripan data dengan rumus *Euclidean Distance* digunakan persamaan berikut [19].

$$d_{(x,y)} = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

b. *Clustering with K-Medoids*

Algoritma K-Medoids juga disebut sebagai algoritma PAM (*Partioning Around Medoids*) yaitu algoritma yang diwakili oleh *cluster* (medoids). Perbedaan algortima K-Means dan K-Medoid terletak pada penentuan perwakilan pusat *cluster* dimana K-Means menggunakan nilai rerata (*mean*) sebagai pusat *cluster* sedangkan K-Medoids menggunakan objek (*medoids*) mewakili pusat *cluster* pada tiap *cluster* [20].

Prosedur algoritma K-Medoids sebagai berikut.

1. Lakukan inisialisasi pusat *cluster* sebanyak jumlah *cluster* (k).
2. Distribusikan setiap objek ke *cluster* terdekat menggunakan *Euclidean Distance* dengan rumus persamaan (5).
3. Seleksi objek secara acak pada tiap-tiap *cluster* sebagai calon *medoid* baru.

 ISSN:2302-8149
e-ISSN:2540-9719

4. Hitung jarak antar objek pada masing-masing *cluster* dengan calon *medoid* baru.
5. Hitung total simpangan (S) dengan menghitung nilai total jarak baru – total jarak lama. Jika didapatkan S < 0, tukarlah objek dengan data *cluster* untuk membuat sekumpulan *k* objek baru sebagai *medoid*.
6. Ulangi langkah 3 sampai dengan 5 hingga tidak terjadi perubahan *medoid*, sehingga diperoleh *cluster* serta anggota cluster masing-masing.

3.4 Uji Validitas

Metode uji validitas menggunakan *Davies-Bouldin Index* yaitu metode yang menghitung rata-rata nilai setiap titik pada himpunan data. Mengetahui hasil *cluster* yang optimal dengan nilai DBI paling kecil atau mendekati nilai (*non-negatif* >= 0) menandakan *cluster* semakin baik [20].

Untuk mencari nilai *Davies Bouldin Index* (DBI) dengan menggunakan persamaan berikut.

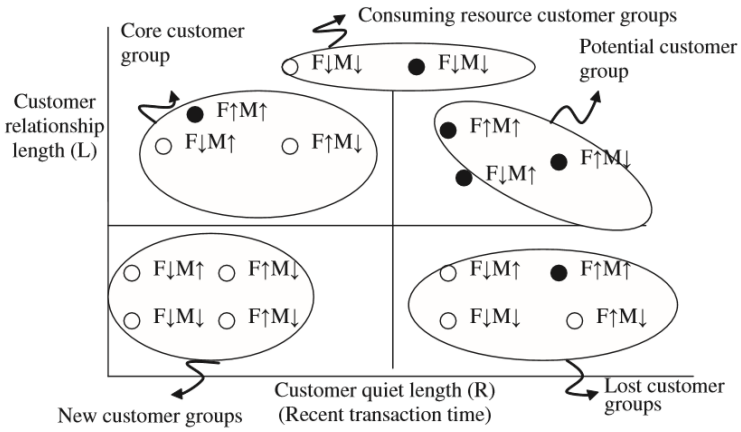
$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\,(R_{i,j}) \quad (6)$$

3.5 Analisis Hasil *Clustering*

Hasil *clustering* pada data dilakukan analisis pada tiap-tiap *cluster*nya menggunakan perhitungan standar deviasi yang mengacu pada model LRFM index. Hasil yang terbentuk dari *cluster* akan menjadi segmentasi yang mewakili karakteristik dan kelompok pelanggan. Bedasarkan model *Length*, *Recency*, *Frequency* dan *Monetary* (LRFM) menjadi model terbaik dari RFM oleh Chang dan Tasy (2004) menambahkan parameter L sebagai variabel untuk mempertimbangkan panjang hubungan pelanggan dengan perusahaan [21]. Berikut penjelasan variabel LRFM [10].

1. *Length* yaitu lamanya hubungan antara pelanggan dan perusahaan yang diukur selama periode analisis.
2. *Recency* yaitu tanggal terakhir transaksi yang dilakukan pelanggan pada perusahaan selama periode analisis.
3. *Frequency* adalah jumlah transaksi yang dilakukan oleh pelanggan pada perusahaan selama periode yang dianalisis.
4. *Monetary* yaitu jumlah uang yang dikeluarkan pelanggan untuk perusahaan selama periode analisis.

Matrik loyalitas pelanggan pada Gambar 2 mengklasifikasi pelanggan menjadi 5 kelompok pelanggan yang terdiri dari 16 jenis kelompok. Adapun makna simbol (↑) merepresentasikan nilai rata-rata *cluster* lebih besar dari nilai rata-rata keseluruhan. Sedangkan simbol (↓) merepresentasikan nilai rata-rata *cluster* lebih kecil dari nilai rata-rata keseluruhan [22].



Gambar 2. *Customer Loyalty Matrix*

Lima kelompok tersebut adalah: (1) *Core Customer (CC)*, terdiri dari pelanggan dengan nilai kesetiaan tinggi (LRFM ↑↓↑↑), pelanggan dengan frekuensi pembelian tinggi (LRFM ↑↓↑↓), dan pelanggan platinum (LRFM ↑↓↓↑); (2) *Potential Customers (PC)* terdiri dari pelanggan berpotensi loyal (LRFM ↑↑↑↑), pelanggan potensial dengan frekuensi yang tinggi (LRFM ↑↑↑↓), dan pelanggan potensial dengan konsumsi (LRFM ↑↓↑↑); (3) *Lost Customers (LC)*, terdiri dari pelanggan bernilai

 ISSN:2302-8149
e-ISSN:2540-9719

tinggi yang hilang dengan (LRFM ↓↑↑↑), pelanggan dengan frekuensi tinggi yang hilang (LRFM ↓↑↑↓), dan pelanggan konsumsi tinggi yang hilang (LRFM ↓↑↓↑), dan pelanggan tidak meyakinkan yang hilang (LRFM ↓↑↓↓); (4) *New Customers (NC)*, terdiri dari nilai tinggi pelanggan baru (LRFM ↓↓↑↑), pelanggan dengan frekuensi promosi tinggi (LRFM ↓↓↑↓), pelanggan dengan pembelanjaan promosi (LRFM ↓↓↓↑), dan pelanggan baru yang tidak menentu (LRFM ↓↓↓↓); (5) *Consuming Resource Customers (CRC)*, terdiri dari pelanggan konsumtif dengan pengeluaran biaya rendah (LRFM ↑↓↓↓), dan pelanggan dengan konsumsi biaya tinggi (LRFM ↑↑↓↓). Hasil perhitungan standar deviasi dari masing-masing atribut dari masing-masing *cluster* akan dibandingkan dengan rata-rata standar deviasi pada masing-masing atribut. Apabila hasil standar deviasi dari *cluster* lebih besar dari rata-ratanya akan disimbolkan dengan tanda panah keatas (↑), sedangkan hasil standar deviasi dari *cluster* lebih rendah dari rata-ratanya disimbolkan dengan panah kebawah (↓) [10].

4 Hasil dan Pembahasan

Dataset online retail terdiri dari delapan atribut *InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID,* dan *Country* dengan jumlah data transaksi sebanyak 541909 *instance*. *Preprocessing* terhadap data dengan melakukan *data cleaning* (Tabel 2) dari empat atribut Adapun atribut yang relevan sesuai model LRFM yaitu atribut *Quantity, InvoiceDate, UnitPrice,* dan *CustomerID.* Pembersihan data dengan nilai yang tidak konsisten/kosong, mengandung nilai minus (-), nilai 0, dan data duplikasi. Hasilnya jumlah data bersih sebanyak 397884 *instance*.

Tabel 2.Data *Cleaning*

Atribut	Keterangan
Quantity	Menghilangkan baris data pada atribut *quantity* dengan nilai minus (-). Pada penelitian ini data dengan nilai tersebut dihapus sejumlah 10624 baris data.
UnitPrice	Menghilangkan baris data pada atribut *unitprice* dengan nilai minus (-) dan nol (0). Pada penelitian ini data dengan nilai tersebut dihapus sejumlah 1181 baris data.
CustomerID	Menghilangkan baris data pada atribut *customerid* dengan nilai nol (0). Pada penelitian ini data dengan nilai tersebut dihapus sejumlah 132220 baris data.

Pertimbangan relevansi data dengan kebutuhan pengujian maka pemilihan atribut *Quantity, InvoiceDate, UnitPrice,* dan *CustomerID* (Gambar 3) sebagai *input* data pengujian dilakukan proses *attribute construction* berdasarkan model LRFM.

CustomerID	Quantity	InvoiceDate	UnitPrice
17850	6	01/12/2010 08:26	2,55
17850	8	01/12/2010 08:26	2,75
17850	6	01/12/2010 08:26	3,39
17850	6	01/12/2010 08:26	3,39
17850	6	01/12/2010 08:26	3,39
17850	6	01/12/2010 08:26	4,25
17850	2	01/12/2010 08:26	7,65

Gambar 3. Atribut Digunakan

Ekstraksi dan transformasi data menyesuaikan model dari atribut LRFM yaitu (*Length*, *Recency*, *Frequency* dan *Monetary*) dapat dilihat pada Gambar 4 dan 5 dengan penjelasan sebaga berikut.

a. *Length* (L)
   Nilai numerik dari pengurangan tanggal antara pembelian terakhir dengan pembelian pertama kali yang dilakukan pelanggan bersumber dari data transaksi pelanggan pada periode penelitian. Nilai *length* pada penelitian ini diperoleh dari menentukan selisih antara tanggal maksimum dan tanggal minimum dari atribut *InvoiceDate.*
b. *Recency* (R)
   Nilai numerik yang didapatkan dari hasil pengurangan tanggal pada transaksi akhir pelanggan dengan tanggal analisis data yang ditentukan peneliti. Nilai *recency* pada penelitian ini diperoleh

 ISSN:2302-8149
e-ISSN:2540-9719

dari menentukan selisih antara tanggal akhir transaksi pelanggan dari tanggal maksimum pada atribut *InvoiceDate.*

c. *Frequancy* (F)
Nilai *frequency* dalam format numerik diperoleh dari jumlah transaksi yang terdapat pada atribut *InvoiceDate*. Nilai *frequency* pada penelitian ini diperoleh dari menentukan jumlah transaksi pada atribut *InvoiceDate* berdasarkan nilai *CustomerID.*

d. *Monetary* (M)
Nilai *monetary* diperoleh dari akumulasi perkalian antara atribut *Quantity* dengan *UnitPrice* yang dilakukan pelanggan pada data transaksi dengan membuat atribut baru Total sebagai penampung nilai akumulasi.

Row Labels	Min of InnvoiceDateConv	Max of InnvoiceDateConv	Count of InnvoiceDateConv	Sum of Total
12346	18/01/2011	18/01/2011	1	77183,6
12347	07/12/2010	07/12/2011	182	4310
12348	16/12/2010	25/09/2011	31	1797,24
12349	21/11/2011	21/11/2011	73	1757,55
12350	02/02/2011	02/02/2011	17	334,4
12352	16/02/2011	03/11/2011	85	2506,04
12353	19/05/2011	19/05/2011	4	89

Gambar 4. Ekstraksi Model LRFM

Tanggal 20 Desember 2011 diacu sebagai periode analisis dalam penelitian ini. Hasilnya diperoleh jumlah pelanggan sebanyak 4338 *instance* berdasarkan *CustomerID*. Hasil dari perhitungan dan ekstraksi model LRFM (Gambar 5) tanpa melakukan pembobotan nilai. Kemudian memastikan data bersih dari kemunculan *outlier* digunakan rumus rentang interkuartil (IQR) yaitu IQR = Q3–Q1. Hasil *outlier detection* ditemukan sebanyak 732 *instance* data yang masuk dalam rentang data *oulier* dan dilakukan pembersihan. Selanjutnya dilakukan normalisasi data.

CustomerID	*Length*	*Recency*	*Frekuency*	*Monetary*
12346	0	336	1	77183,6
12347	365	13	182	4310
12348	283	86	31	1797,24
12349	0	29	73	1757,55
12350	0	321	17	334,4
12352	260	47	85	2506,04

Gambar 5. Hasil Perhitungan Model LRFM

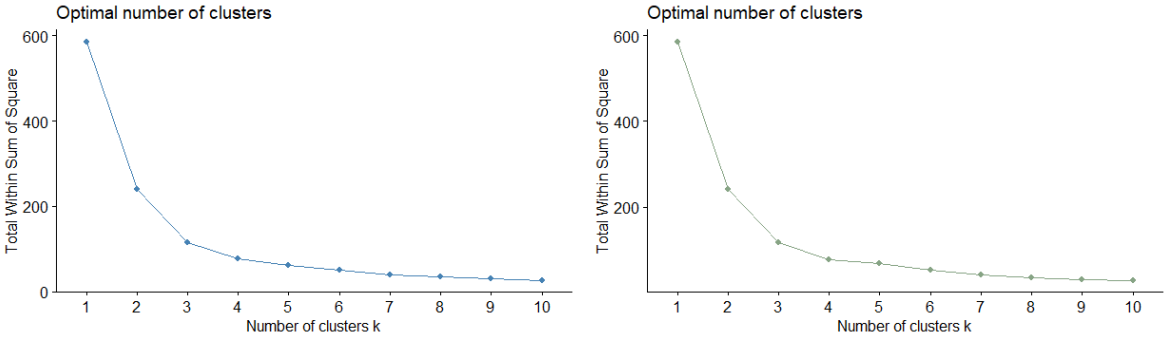
Normalisasi data menggunakan metode min-max normalisasi menjadi rentang nilai antara 0 dan 1 yang hasilnya dapat dilihat pada Gambar 6. Sehingga jumlah data pelanggan unik berjumlah 4338 *instance* data berubah menjadi 3606 *instance* data. Langkah selanjutnya dari ke 3606 *instance* data yang merupakan hasil akhir dari tahapan *preprocessing* dan dilanjut ke proses penentuan *cluster* optimal.

CustomerID	*LN*	*RN*	*FN*	*MN*
12348	0,76	0,20	0,00	0,01
12349	0,00	0,05	0,01	0,01
12350	0,00	0,83	0,00	0,00
12352	0,70	0,10	0,01	0,01
12353	0,00	0,55	0,00	0,00
12354	0,00	0,62	0,01	0,00

Gambar 6. Data Hasil Normalisasi

 ISSN:2302-8149
e-ISSN:2540-9719

Pada tahapan *clustering* dengan algoritma K-Means dan K-Medoids dilakukan pengujian pada 3606 *instance* data. Penentuan nilai *cluster k* yang optimal digunakan metode *Elbow* dengan program R Studio., dimulai dari *k = 2* dan seterusnya dan hasil sebagaimana Gambar 7 berikut.



K-Means K-Medoids

Gambar 7. Visualisasi Grafik Nilai *k* Menggunakan Metode Elbow

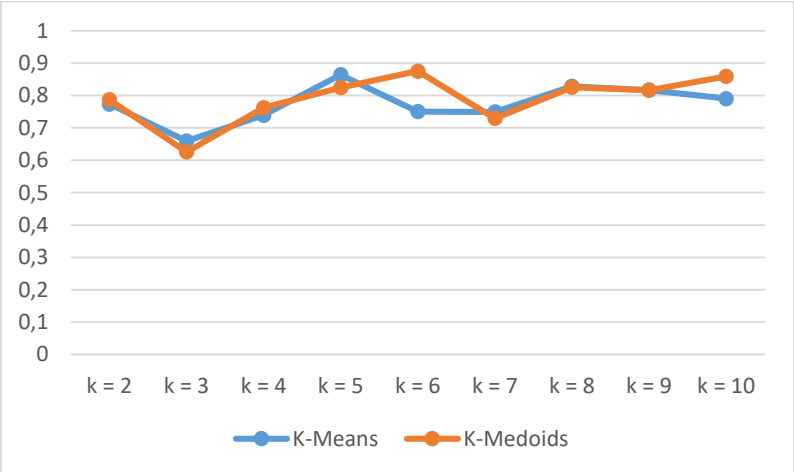
Berdasarkan visualisasi grafik pada Gambar 7, maka disimpulkan nilai *cluster k* optimal dari metode *elbow* tersebut adalah nilai *k=3* sebagai nilai *k* yang optimal dari masing-masing pengujian menggunakan algoritma K-Means dan K-Medoids. Pada masing-masing grafik juga menunjukan tidak terdapat perubahan nilai yang signifikan mulai dari *cluster k* = 4 dan seterusnya.

Pengujian dilanjutkan dengan metode *Davies Bouldin Index* (DBI) untuk mengevaluasi *cluster* terbaik. Pengujian dilakukan dari *cluster k* = 2 sampai *k* = 10. Hasil pengujian dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian DBI

Cluster	**K-Means**	**K-Medoids**
k = 2	0,7736318	0,7868374
k = 3	**0,659677**	**0,6256833**
k = 4	0,7391369	0,7623837
k = 5	0,8641701	0,8244547
k = 6	0,7504408	0,8756588
k = 7	0,7495017	0,7297279
k = 8	0,8286562	0,8255569
k = 9	0,8165835	0,81671
k = 10	0,7903605	0,8594031

Berdasarkan hasil pengujian nilai DBI diatas, memperjelas bahwa *k* = 3 sebagai *cluster* optimal dari masing-masing algoritma, perhatikan grafik pada Gambar 8.



Gambar 8. Grafik DBI K-Means Dan K-Medoids

 ISSN:2302-8149
e-ISSN:2540-9719

Untuk mengevaluasi *cluster* yang terbentuk dapat menggunakan nilai *average within* dan *average between cluster* sebagai acuan penujian. *Cluster* disebut terbaik jika mempunyai nilai *average within* yang sangat kecil dan memiliki *average between* yang sangat besar sebagaimana Tabel 4.

Tabel 4. Nilai Average Within Dan Average Between K-Means Dan K-Medoid

	K-Means	**K-Medoids**
average within	0,2171401	0,2169299
average between	0,6422888	0,6426125

Berdasarkan Tabel 4 diatas disimpulkan bahwa nilai *average within* K-Means lebih kecil dari nilai *average between* K-Means, begitu juga dengan K-Medoids. Sehingga disimpulkan bahwa *cluster* yang terbentuk dari K-Means dan K-Medoids sudah baik dan optimal.

Kemudian menentukan algoritma yang performa baik dengan melihat nilai rationya. *Cluster* dengan ration paling kecil adalah *cluster* terbaik. Nilai ration ini diperoleh dari hasil pembagian *average within* dengan *average between*. Dari hasil perhitungan diperoleh bahwa ration K-Means sebesar 0,3380724 sedangkan ration K-Medoids sebesar 0,337575 dan dapat disimpulkan bahwa ration yang paling kecil dari K-Medoids.

Berdasarkan hasil pengujian nilai *cluster k* optimal adalah *k* = 3, maka *clustering* pada data transaksi menggunakan algoritma terbaik yaitu K-Medoids dengan *cluster k* =3. Pada tahapan analisis hasil *clustering* bertujuan untuk mengetahui kelompok pelanggan termasuk dalam karakteristik dan kelompok pelanggan per segmen sesuai jumlah *cluster* yang terbentuk. Hasil analisis ini menjadi acuan bagi perusahaan dalam menyusun strategi pemasaran dan informasi mengenai karakteristik pelanggannya. Hasil yang terbentuk dari proses *clustering* K-Medoids diperoleh 3 (tiga) *cluster* / kelompok pelanggan dari 3606 *instance* yaitu *cluster* 1 sebanyak 1378 *instance*, *cluster* 2 sebanyak 1373 *instance,* dan *cluster* 3 sebanyak 855 *instance.* Hasil 3 (tiga) kelompok data pelanggan tersebut dilakukan proses perhitungan standar deviasi tiap-tiap atribut L, R, F, dan M dari masing-masing *cluster*. Hasilnya diklasifikasikan berdasarkan nilai LRFM index. Klasifikasian menggunakan simbol (↑) untuk standar deviasi yang tinggi dan simbol (↓) untuk standar deviasi yang rendah dari rerata keseluruhan. Hasil perhitungan standar deviasi dan rata-rata dari ketiga *cluster* sebagai Tabel 5 berikut.

Tabel 5. Nilai Standar Deviasi Model LRFM

Cluster	**Jumlah Pelanggan**	**L**		**R**		**F**		**M**	
1	1378	0,177670394	**L↑**	0,106810161	**R↓**	0,006617113	**F↑**	0,003095601	**M↑**
2	1373	0,10592539	**L↓**	0,102215247	**R↓**	0,004862968	**F↓**	0,002050805	**M↓**
3	855	0,114169527	**L↓**	0,136373012	**R↑**	0,003361743	**F↓**	0,001633505	**M↓**
Rata-rata	**3606**	**0,136000152**		**0,113848539**		**0,005008221**		**0,002285883**	

Berdasarkan matrix loyalitas pelanggan ditunjukkan pada Gambar 2, klasifikasi berdasarkan LRFM index dengan representasi simbol memberikan penjabaran antara *customer loyalty matrix* yang terdiri dari 16 grup pelanggan dengan hasil *clustering* diperoleh hasil 3 ketiga kelompok pelanggan termasuk dalam beberapa kelompok berikut.

a. *Cluster* 1 sebagai kelompok pelanggan pelanggan *Core Customer* (CC) yaitu grup pelanggan *Including High Value Loyal Customers* (*length* ↑, *recency* ↓, *frequency* ↑, dan *monetary* ↑) terdiri dari 1378 pelanggan
b. *Cluster* 2 sebagai kelompok pelanggan *New Customer* (NC) yaitu grup pelanggan *Uncertain New Customer* (*length* ↓, *recency* ↓, *frequency* ↓, *dan monetary* ↓) terdiri dari 1373 pelanggan.
c. *Cluster* 3 sebagai kelompok pelanggan *Lost Customer* (LC) yaitu grup pelanggan *Uncertain Lost Customers* (*length* ↓, *recency* ↑, *frequency* ↓, *dan monetary* ↓) terdiri dari 855 pelanggan

 ISSN:2302-8149
e-ISSN:2540-9719

5 Kesimpulan

Penerapan algoritma K-Means dan K-Medoids pada data transaksi *e-commerce* untuk menghasilkan segmentasi pelanggan. Pada penelitian ini algoritma terbaik ditunjukan oleh algoritma K-Medoids sesuai hasil perhitungan nilai ration dari masing-masing algoritma yaitu K-Means sebesar 0,3380724 sedangkan ration K-Medoids sebesar 0,337575. Pada penentuan nilai *cluster* optimal, algoritma K-Means dan K-Medoids sama-sama memberikan nilai *cluster k = 3*. Sehingga pada penelitian ini digunakan algoritma K-Medoids dan *cluster* optimal *k* = 3 sesuai dengan hasil metode *Elbow* dan uji validitas *Davies-Bouldin Index*. Hasil analisis segmentasi pelanggan *e-commerce* berdasarkan *customer loyalty matrix* terdapat 3 segmen / kelompok pelanggan yaitu *Core Customer, New Customers*, dan *Lost Customer*. Masing-masing pelanggan memiliki karakteristik dan perilaku yang berbeda-beda, sehingga perusahan saat ini dapat dengan mudah mengenali setiap pelanggan berdasarkan segmentasi pelanggan. Perusahaan bisa mengenali kelompok pelanggan berdasarkan hasil *clustering* sebagai langkah dalam menciptakan strategi *marketing*, dan rekomendasi upaya dalam menjalin hubungan antara pelanggan dan perusahaan

6 Referensi

[1] J. Celement, 29 Oktober 2020. [Online]. Available: https://www.statista.com/topics/871/online-shopping.

[2] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, F. and T. M. A. Aziz, "Segmentation Model of Customer Lifetime Value in Small an Medium Enterprise (SMEs) using K-Means Clustering and LRFM Model," *International Journal of Integrated Engineering*, vol. 11, pp. 169 -180, 2019.

[3] O. Dogan, E. Aycin and Z. A. Bulut, "Customer Segmentation by Using RFM Model and Clustering Method: A Case Study in Retail Industry," *International Jurnal of Contemporary Economics and Administrative Sciences*, vol. 8, pp. 1-19, 2018.

[4] B. Kaur and P. K. Sharma, "Implementation of Customer Segmentation using Integrated Approach," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6S, pp. 770 - 772, 2019.

[5] H. Singh and S. Srivastava, "Customer Segmentation in E-Commerce to Retain and Gain the Customers," *International Journal of Advanced Science and Technology*, vol. 29, no. 7, pp. 12846-12856, 2020.

[6] D. Kandeil, A. Saad and S. M. Youssef, "A Two-phase Clustering Analysis for B2B Customer Segmentation," *International Conference on Intelegent Networking and Collaborative System*, pp. 221 - 228, 2014.

[7] V. Babaiyan and S. A. Sarfarazi, "Analyzing Customers of South Khorasan Telecommunication Company with Expansion of RFM to LRFM Model," *Journal of AI and Data Mining*, vol. 7, no. 2, pp. 331 - 340, 2019.

[8] E. U. Wahyuningtyas, R. R. M. Putri and S. , "Optimasi K-Means Untuk Clustering Dosen Berdasarkan Kinerja Akademik Menggunakan Algoritme Genetika Paralel," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2628 - 2635, 2018.

[9] C. D. Rumiarti and I. Budi, "Segmentasi Pelanggan Pada Customer Relationship Management di Perusahaan Ritel: Studi Kasus PT Gramedia Asri Media," *Jurnal Sistem Informasi (Jurnal of Information System)*, vol. 13, no. 1, pp. 1 - 10, 2017.

[10] S. Monalisa, "Klasterisasi Customer Lifetime value dengan Model LRFM Menggunakan Algoritma K-Means," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 5, pp. 247 - 252, 2018.

[11] R. Gustriansyah, N. Suhandi and F. Antony, "Clustering Optimization in RFM Analysis Based on K-Means," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470 - 477, 2020.

[12] M. Aryuni, E. D. Madyatmadja and E. Miranda, "Penerapan K-Means Dan K-Medoids Clustering Pada Data Internet Banking Di Bank XYZ," *Jurnal Teknik dan Ilmu Komputer*, vol. 7, no. 27, pp. 349 - 356, 2018.

ISSN:2302-8149
e-ISSN:2540-9719

[13] D. Chen, L. S. Sain and K. Guo, "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*, vol. 19, no. 3, pp. 197 - 208, 2012.

[14] J. Han, M. Kamber and J. Pie, Data mining: concepts and techniques, 3rd ed., United States of America: Morgan Kaufmann Publishers is an imprint of Elsevier, 2012.

[15] M. Hubert and S. V. d. Veeken, "Outlier Detection for Skewed Data," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 22, no. 3-4, pp. 235 - 246, 2008.

[16] H. W. Alomari and M. Stephan, "Towards slice-based semantic clone detection," in *2018 IEEE 12th International Workshop on Software Clones (IWSC)*, Campobasso, Italy, 2018.

[17] S. Adinugroho and Y. A. Sari, Implementasi Data Mining Menggunakan Weka, Malang: Universitas Brawijaya Press, 2018.

[18] J. Qi, Y. Yu, L. Wang and J. Liu, "K*-Means: An Effective and Efficient K-Means Clustering Algorithm," Atlanta, GA, 2016.

[19] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski DIstance, dan Manhattan DIstance pada Algoritma K-Means Clustering berbasis Chi-square," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol. 4, no. 1, pp. 20 - 24, 2019.

[20] I. Kamila, U. Khairunnisa and M. , "Perbandingan Algoritma K-Means dan K-Medoids untuk PengelompokanData Transaksi Bongkar Muat di Provinsi Riau," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 5, no. 1, pp. 119 - 125, 2019.

[21] R. A. Daoud, A. Amine, B. Bouikhalene and R. Lbibb, "Customer Segmentation Model in E-Commerce Using Clustering Techniques and LRFM Model: The Case of Online Store in Morocco," *International Journal of Computer and Information Engineering*, vol. 9, no. 8, pp. 2000 - 2010, 2015.

[22] D. C. Li, W. L. Dai and W. T. Tseng, "A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business," *Expert System with Applications*, vol. 38, no. 6, pp. 7186 - 7191, 2011.

[23] G. N. W. Paramartha, D. E. Ratnawati and A. W. Widodo, "Analisis Perbandingan Metode K-Means Dengan Improved Semi-Supervised K-Means Pada Data Indeks Pembangunan Manusia (IPM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1, no. 9, pp. 813 - 824, 2017.

[24] E. Irwansyah and M. Faisal, Advanced Clustering: Teori dan Aplikasi, Yogyakarta: DeePublish, 2015.