ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 3, Tahun 2022, Halaman 429 - 438
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

IMPLEMENTASI *K-MEDOIDS* DAN MODEL *WEIGHTED-LENGTH RECENCY FREQUENCY MONETARY* (W-LRFM) UNTUK SEGMENTASI PELANGGAN DILENGKAPI GUI R

Ta'fif Lukman Afandi[1*], Budi Warsito[2], Rukun Santosa[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***email**:tafif5880@gmail.com

DOI: 10.14710/J.GAUSS.11.3.429-438

Article Info:
Received: 2022-08-03
Accepted: 2022-10-15
Available Online: 2023-01-03

Keywords:
K-Medoids, W-LRFM Model, Customer Segmentation, GUI R.

Abstract: The k-medoids algorithm is a partition-based clustering algorithm that groups n objects as much as k clusters. The algorithm uses medoids as the center point (partition) of the cluster. Medoids are actual objects that are randomly selected as the most centered object in a cluster so that the k-medoids algorithm is robust against outliers. Grouping objects in cluster analysis based on similarities between objects. Measurement of similarity between objects can use the euclidean and manhattan distances. The use of distance in cluster analysis can affect cluster results. Validation of cluster results using internal validation, namely the silhouette index. The Weighted-Length Recency Frequency Monetary (W-LRFM) model is a model that applies the relative importance (weight) of the LRFM model according to the importance of each variable in the LRFM model. LRFM model is a model used for customer segmentation based on customer behavior which consists of variables length, recency, frequency, and monetary. The relative importance (weight) of the W-LFRM model uses the Analytics Hierarchical Process (AHP) method. The W-LRFM model is used to calculate the Customer Lifetime Value (CLV) of each cluster. The implementation of k-medoids and the W-LFRM model in this study are used for customer segmentation based on the length, recency frequency, and monetary variable. The formation of these variables is the result of transformation of customer behavior data such as transaction id, date of purchase, and a total amount of 41,073 rows into variable length, recency, frequency, and monetary as much as 5,108 rows. The criteria of the best cluster formed are k = 2 using the manhattan distance with the average of coefficient values = 0.62. The weights on the W-LRFM model produced based on the AHP method are 0.16, 0.29, 0.47, and 0.08 for the variable length, recency, frequency, and monetary. CLV formed from two clusters, namely 0.158 and 0.499. CLV in the second cluster is bigger so that the second cluster becomes the main priority in the marketing strategy. The second cluster has the characteristics 0.29, 0.47, and 0.08 for the variable length, recency, frequency, and monetary. The second cluster has the characteristics $L \uparrow R \downarrow F \uparrow M \uparrow$ means a loyal customer group. The first *cluster* has characteristics $L \downarrow R \uparrow F \downarrow M \downarrow$ means a potential customer group. This research is assisted by using Graphical User Interface (GUI) R to facilitate analysis.

1. PENDAHULUAN

Segmentasi pelanggan merupakan salah satu strategi yang dapat dilakukan untuk membantu perusahaan dalam mengidentifikasi karakterisitik para pelanggan sehingga perusahaan dapat membuat program yang sesuai dengan kebutuhan pelanggan. Salah satu metode statistika yang dapat digunakan untuk segmentasi yaitu analisis *cluster* dengan menggunakan data transaksi berdasarkan perilaku pelanggan. Analisis *cluster* adalah salah satu metode untuk mengelompokkan *instance* (*sample*) menjadi beberapa *group* atau subset atau *cluster* berdasarkan "kemiripan" dengan *instance* lainnya (Primartha, 2021).

K-Medoids merupakan salah satu metode analisis *cluster* yang menggunakan teknik berbasis objek *representative* (perwakilan) yang disebut *medoids* (Suyanto, 2017). *K-Medoids* memiliki beberapa kelebihan antara lain tidak sensitif dengan *outlier*, dapat mengurangi *noise*, dan lebih baik dalam waktu eksekusi (Arora dkk., 2016). Metode *k-medoids* juga memiliki beberapa kelemahan, salah satunya adalah tidak dapat menentukan *cluster* optimal. Penentuan *cluster* optimal pada penelitian ini menggunakan metode validasi *silhouette index*. Hasil *cluster* optimal menggunakan *k-medoids* belum dapat mengurutkan prioritas *cluster* pelanggan hasil pemodelan.

Customer Lifetime Value (CLV) merupakan statistik yang digunakan untuk mengurutkan prioritas *cluster* pelanggan dari pengelompokkan pelanggan. Model *Weighted-Length Recency Frequency Monetary* (W-LRFM) dapat digunakan menghitung CLV. Model tersebut merupakan model yang menerapkan kepentingan relatif (bobot) parameter pada model LRFM sesuai dengan kepentingan masing-masing parameter. Model LRFM merupakan model yang berisi empat parameter (variabel) yaitu *length, recency, frequency,* dan *monetary* sebagai parameter yang menggambarkan perilaku pelanggan. Semakin tinggi CLV pada suatu *cluster* pelanggan maka semakin prioritas *cluster* pelanggan tersebut dalam strategi pemasaran. Bobot dalam model W-LRFM dihitung dengan menggunakan metode *Analytics Hierarchical Process* (AHP). Penelitian ini dalam analisis data menggunakan aplikasi *Graphical User Interface* (GUI) R. Penggunaan aplikasi GUI R bertujuan untuk memudahkan analisis data dan menampilkan *output* yang lebih menarik sehingga memudahkan perusahaan untuk mendapatkan *insight*.

2. TINJAUAN PUSTAKA

Model Weighted-LRFM (W-LRFM) merupakan pengembangan model dari model Weighted-RFM (W-RFM) dengan adanya penambahan satu variabel yaitu *length*. Penambahan variabel *length* digunakan untuk membedakan pelanggan yang memiliki hubungan jangka panjang atau jangka pendek dengan perusahaan (Wei dkk., 2012). Model W-RFM merupakan suatu model yang terdiri dari tiga parameter (*recency, frequency,* dan *monetary*) dengan menerapkan bobot yang berbeda pada masing-masing parameter berdasarkan tingkat kepentingan parameter bagi perusahaan (Carneiro dan Vera, 2021). Bobot model W-LRFM dihitung menggunakan metode *Analytics Hierarchical Process* (AHP). Model W-LRFM pada penelitian ini digunakan untuk menghitung *Customer Lifetime Value* (CLV).

Model W-LRFM memilik empat parameter yaitu *length, recency, frequency,* dan *monetary*. *Length* yaitu parameter yang mengukur pelanggan berdasarkan rentang waktu (tanggal, bulan, tahun) transaksi awal sampai dengan transaksi akhir. *Recency* yaitu parameter yang mengukur pelanggan berdasarkan rentang waktu (tanggal, bulan, tahun) transaksi akhir sampai dengan saat ini (periode yang ditentukan oleh peneliti). *Frequency* yaitu parameter yang mengukur pelanggan berdasarkan jumlah transkasi yang dilakukan setiap pelanggan sampai dengan transaksi terakhir. *Monetary* yaitu parameter yang mengukur pelanggan berdasarkan jumlah uang yang dikeluarkan setiap pelanggan sampai dengan transaksi terakhir.

Salah satu proses dalam *data mining* yaitu data *preprocessing*. Penelitian ini melakukan tiga tahap data preprocessing yaitu transformasi data, deteksi *outlier*, dan normalisasi data. Transformasi data dilakukan untuk merubah struktur data dari data transaksi harian menjadi empat variabel yaitu *length, receny, frequency,* dan *monetary* sebagai variabel yang menggambarkan perilaku pelanggan.

Deteksi *outlier* merupakan salah satu bagian dalam data *preprocessing* dengan mencari keterdapatan data *outlier*. Metode pengukuran jarak kuadrat mahanalobis digunakan untuk

deteksi *outlier* pada penelitian ini. Hal tersebut dikarenakan variabel yang digunakan lebih dari dua sehingga penelitian ini termasuk kasus *multivariate*. *Outlier* pada dimensi data yang tinggi tidak dapat dideteksi melalui plot *univariate* atau *bivariate* (Johson dan Wichern, 2007). Jarak mahanalobis merupakan jarak antar dua objek yang dinyatakan dalam bentuk vektor dan matrik dengan melibatkan kovarians atau korelasi antar peubah. Persamaan untuk menghitung jarak kuadrat mahanalobis untuk setiap objek tercantum pada persamaan (1):

$$d^2_{MD(i)} = (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}), i = 1,2, \dots, o \quad (1)$$

dengan $d^2_{MD(i)}$ merupakan jarak kuadrat mahanalobis objek ke-*i*, x_i merupakan vektor data objek ke-*i* yang berukuran *p x* 1, $\bar{x}$ merupakan vektor rata-rata dari setiap variabel berukuran *p x* 1, dan Σ merupakan matriks kovarian berukuran *p x p*. Menurut Johnson dan Wichern (2007), objek ke-*i* dikatakan *outlier* jika $d^2_{MD(i)} > \chi^2_{p(1-\alpha)}$ dengan $\chi^2_{p(1-\alpha)}$ merupakan nilai *chi-square* berderajat bebas *p* dengan probabilitas $(1-\alpha)$, p merupakan banyaknya variabel, dan $1-\alpha$ merupakan tingkat kepercayaan.

Normalisasi data atau disebut juga standarisasi data merupakan salah satu bagian dalam data *preprocessing* dengan merubah rentang data pada setiap variabel yang bertujuan untuk menghindari bias pada model. Salah satu metode normalisasi yaitu normalisasi *min-max*. Normalisasi *min-max* merupakan metode normalisasi dengan menggunakan nilai maksimum dan minimum untuk mengkonversi data secara linier (Suyanto, 2017). Nilai baru $\hat{x}_{ij}$ yang digunakan pada penelitian ini berada dalam rentang $[0,1]$ dengan persamaan (2):

$$\hat{x}_{ij} = \frac{x_{ij} - min_j}{max_j - min_j}, i = 1,2, \dots, o \text{ dan } j = 1,2, \dots, p \quad (2)$$

dengan $\hat{x}_{ij}$ merupakan nilai hasil normalisasi objek ke-*i* pada variabel ke-*j*, x_{ij} merupakan nilai awal objek ke-*i* pada variabel ke-*j*, min_A merupakan nilai awal terkecil, dan max_A merupakan nilai awal terbesar.

Analisis *cluster* adalah salah satu metode untuk mengelompokkan *instance* (*sample*) menjadi beberapa *group* atau *cluster* berdasarkan kemiripan dengan *instance* yang lain (Primartha, 2021). Analisis *cluster* harus memenuhi dua asumsi sebagai berikut (Nugroho, 2008):

1. Sampel Mewakili Populasi
   Sampel yang digunakan pada analisis *cluster* harus dapat mewakili populasi yang ada. Asumsi sampel mewakili populasi dapat menggunakan uji KMO (*Kaiser Mayer Olkin*). Sampel akan dikatakan layak untuk mewakili populasi apabila nilai KMO > 0,5.
2. Non Multikolinearitas
   Parameter atau variabel yang digunakan pada analisis *cluster* sebaiknya variabel-variabel tidak terindikasi adanya multikolinearitas (Hair dkk., 2010). Variabel terindikasi adanya multikolinearitas jika nilai $VIF > 10$.

Penelitian ini menggunakan dua ukuran jarak untuk membandingkan hasil *cluster* yaitu jarak *euclidean* dan *manhattan*. Jarak *euclidean* merupakan akar jumlah kuadrat perbedaan nilai untuk tiap variabel (Johson dan Wichern, 2007). Persamaan untuk menghitung jarak *euclidean* tercantum pada persamaan (3):

$$d_{euc}(\boldsymbol{i}, \boldsymbol{h}) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{kj}\right)^2}, i = 1,2, \dots, o, k = 1,2, \dots, o, \text{ dan } j = 1,2, \dots, p \quad (3)$$

dengan x_{ij} merupakan data dari objek ke-*i* pada variable ke-*j* dan x_{hj} merupakan data dari objek ke-*h* pada variable ke-*j*. Jarak *manhattan* disebut juga jarak *city blok* merupakan jumlah perbedaan mutlak tiap variabel (Johson dan Wichern, 2007). Persamaan untuk menghitung jarak *manhattan* tercantum pada persamaan (4):

$$d_{cb}(\boldsymbol{i}, \boldsymbol{h}) = \sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right|, i = 1,2, \dots, o, k = 1,2, \dots, o, \text{ dan } j = 1,2, \dots, p \quad (4)$$

dengan x_{ij} merupakan data dari objek ke-*i* pada variable ke-*j* dan x_{hj} merupakan data dari objek ke-*h* pada variable ke-*j*.

Algoritma yang digunakan adalah algoritma *k-medoids*. Algoritma *k-medoids* atau dikenal sebagai *Partitioning Around Medians* (PAM) merupakan salah satu algoritma *clustering* berbasis partisi yang melakukan pengelompokan objek ke dalam *k cluster* berdasarkan *medoids* sebagai pusat klaster (Santoso dan Umam, 2018). *Medoids* adalah nilai yang diambil dari salah satu datanya itu sendiri yang dapat memberikan nilai rerata ketidakmiripan (*dissimilarity*) paling kecil. Hal tersebut yang menjadikan algoritma *k-medoids* lebih *robust* terhadap *noise* dan *outlier*. Tahapan-tahapan algoritma *k-medoids* sebagai berikut (Suyanto, 2017):

1. Menentukan *k* sebagai jumlah *cluster* yang akan dibentuk
2. Membangkitkan *k* pusat *cluster* (*medoids*) secara acak sebagai objek-objek *representative* awal
3. Menghitung jarak objek non *representative* (*non-medoids*) dengan objek *representative* (*medoids*) tiap *cluster* dan memasukan tiap objek non *representative* (*non-medoids*) ke dalam *cluster* yang memiliki objek *representative* (*medoids*) terdekat, kemudian hitung total jaraknya (E_a) dengan persamaan (5):
$$E_a = \sum_{i=1}^{o} D_i \ , i = 1,2, \dots, o \quad (5)$$
dengan D_i merupakan minimum jarak antara objek ke-*i* dan *medoids* awal di masing-masing *cluster*.
4. Memilih secara acak objek non *representative* (*non-medoids*) pada masing-masing *cluster* sebagai kandidat *medoids* baru yang dinotasikan sebagai $h_{t+1,k}$
5. Menghitung jarak objek non *representative* (*non-medoids*) dengan kandidat *medoids* baru dan memasukan tiap objek non *representative* (*non-medoids*) ke dalam *cluster* yang memiliki objek *representative* (*medoids*) terdekat, kemudian hitung total jaraknya (E_{random}) dengan persamaan (6):
$$E_{random} = \sum_{i=1}^{o} D_{random(i)} \ , i = 1,2, \dots, o \quad (6)$$
dengan $D_{random(i)}$ merupakan minimum jarak antara objek ke-*i* dan *medoids* baru di masing-masing *cluster*.
6. Menghitung selisih total jarak (S) dengan persamaan (7):
$$S = E_{random} - E_a \quad (7)$$
7. Jika selisih total jarak $(S) < 0$, maka E_{random} menggantikan E_a sebagai objek *representative* (*medoids*) baru dan jika selisih total jarak $(S) > 0$ iterasi berhenti.
8. Mengulangi langkah keempat sampai dengan langkah ketujuh sampai konvergen yaitu selisih total jarak $(S) > 0$.

Algoritma *k-medoids* memiliki kelemahan salah satunya tidak dapat menentukan partisi yang optimal. Salah satu metode untuk membantu algoritma *k-medoids* menentukan partisi optimal yait *silhouette coefficient*. *Silhouette Coefficient* merupakan sebuah metode yang digunakan untuk validasi hasil *cluster* pada data berskala rasio dengan menggabungkan metode separasi dan metode kohesi (Kauffman dan Rousseeuw, 1990). Persamaan untuk menghitung rata-rata nilai koefisien *silhouette* seperti pada persamaan (8):

$$SC = \frac{1}{o}\sum_{i=1}^{o} s(x_i) \, , i = 1,2, \dots, o \quad (8)$$

dengan $s(x_i) = \frac{b_i - a_i}{max\{b_i, a_i\}}$, a_i merupakan rata-rata kemiripan objek *i* dengan lainnya dalam satu *cluster*, dan b_i merupakan minimum rata-rata kemiripan objek *i* dengan lainnya di masing-masing *cluster*. Kriteria subjektif kualitas pengelompokan berdasarkan rata-rata

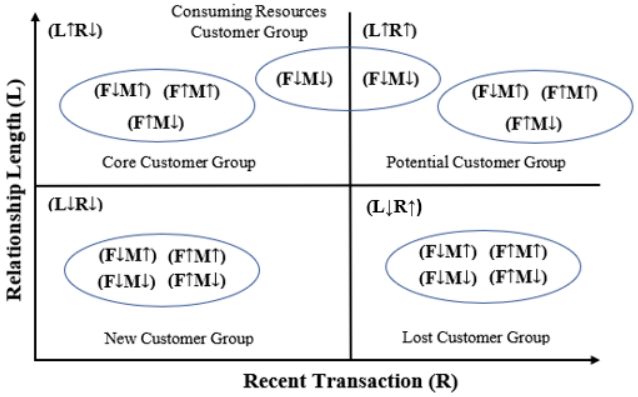
nilai koefisien *silhouette* dapat dikelompokkan menjadi empat yaitu rata-rata nilai koefisien *silhouette* antara 0,00 – 0,25 diintrepretasikan *bad cluster*, rata-rata nilai koefisien *silhouette* antara 0,26 – 0,50 diintrepretasikan *weak cluster*, rata-rata nilai koefisien *silhouette* antara 0,51 – 0,70 diintrepretasikan *good cluster*, dan rata-rata nilai koefisien *silhouette* antara 0,71 – 1,00 diintrepretasikan *strong cluster* (Kauffman dan Rousseeuw, 1990).

Pengelompokan pelanggan pada penelitian ini berdasarkan empat variabel yaitu *length*, *recency*, *frequency*, dan *monetary*. Hasil pengelompokan pelanggan berdasarkan *cluster* optimal belum dapat menentukan prioritas kelompok secara sendiri. Penentuan prioritas kelompok pelanggan dapat dibantu dengan menghitung *Customer Lifetime Value* (CLV). Perhitungan CLV ini menggunakan model W-LRFM. CLV dapat diestimasi secara individu setiap pelanggan maupun setiap segmen pelanggan atau *cohort* (Stan dan Buttle, 2015). Formula untuk memprediksi CLV dalam *cluster* tercantum pada persamaan (9):

$$CLV = C_L^J.W_L + C_R^J.W_R + C_F^J.W_F + C_M^J.W_M \quad (9)$$

dengan $C_L^J, C_R^J, C_F^J, C_M^J$ merupakan nilai rata-rata *cluster* dari setiap parameter dan W_L, W_R, W_F, W_M merupakan bobot dari setiap parameter model W-LRFM dari hasil *Analytics Hierarchical Process* (AHP).

Langkah terakhir dalam penelitian ini yaitu interpretasi *cluster*. Interpretasi *cluster* digunakan untuk untuk memberi nama spesifik dan untuk mengetahui profil dari setiap *cluster*. Pemberian nama dalam segmentasi model LRFM telah dikelompokkan oleh Chang dan Tsay yang berawal 16 kelompok menjadi lima kelompok (Kandeil dkk., 2014). Simbol (↑) didefinisikan nilai rata-rata variabel dalam *cluster* di atas rata-rata keseluruhan objek pada variabel tersebut sedangkan simbol (↓) didefinisikan nilai rata-rata variabel dalam *cluster* di bawah rata-rata keseluruhan objek pada variabel tersebut.



Gambar 1. Matriks Kesetiaan Pelanggan

Kompleksitas algoritma *k-medoids* yang tinggi sehingga perlu adanya alat bantu salah satunya *software* R. *Software* R memiliki dua jenis *user interface* dalam *software* R yaitu *Command Line Interface* (CLI) dan *Graphical User Interface* (GUI). Penelitian ini *user interface* yang digunakan yaitu GUI. Paket R yang digunakan untuk membuat GUI R yaitu *shiny*. Secara umum, komponen R *Shiny* dibedakan menjadi dua kelompok besar, yaitu *User Interface* (UI) dan *server* (Tirta, 2014). *User Interface* (UI) merupakan komponen yang mendefinisikan tampilan web dari aplikasi yang memuat seluruh *input* dan *output* sedangkan *server* merupakan otak dari program yang bertugas melakukan simulasi, bermacam analisis data sesuai dengan pilihan pengguna yang selanjutnya hasilnya dikirim ke bagian *output*.

3. METODE PENELITIAN

Jenis data yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder yang digunakan meminta data transaksi harian pada salah satu perusahaan farmasi di Jakarta. Data tersebut berupa data histori transaksi pelanggan salah satu *brand* perusahaan farmasi di Jakarta dengan mengambil dari tiga grup *channel* yang dinotasikan sebagai A, B, dan C pada periode 2021 serta pelanggan berdomisili di daerah Jakarta, Bogor, Depok, Tangerang, dan Bekasi. Variabel yang digunakan pada penelitian ini yaitu Kontak ID, ID Transaksi, *Purchased Date*, dan Total *Amount* dengan jumlah 41.073 *row* data. Variabel tersebut diolah menggunakan MySQL dan *Graphical User Interface* (GUI) R dari *package Shiny*. Langkah-langkah analisis data dalam penelitian ini diuraikan sebagai berikut:

1. Mengumpulkan data transaksi harian.
2. Melakukan transformasi data dengan merubah data transaksi harian menjadi data LRFM.
3. Melakukan analisis *cluster* menggunakan GUI R. Langkah-langkah analisis *cluster* sebagai berikut:
    a. Memasukan data hasil transformasi menggunakan *tools* MySQL ke GUI R.
    b. Mengecek keterdapatan data *outlier* menggunakan pengukuran jarak kuadrat mahanalobis.
    c. Melakukan normalisasi data menggunakan metode *min-max*.
    d. Melakukan uji asumsi *cluster* yaitu menguji sampel yang mewakili dan uji non multikolinearitas.
    e. Mencari *cluster* optimal dengan membandingkan kriteria *cluster* yaitu ukuran jarak *euclidean* dan *manhattan* serta batas k *cluster* yang akan diuji.
    f. Melakukan pemodelan *cluster* optimal.
    g. Menghitung bobot model W-LRFM menggunakan metode AHP.
    h. Menghitung CLV berdasarkan model W-LRFM.
    i. Melakukan interpretasi dan profilisasi *cluster*.

4. HASIL DAN PEMBAHASAN

Segmetasi pelanggan dalam penelitian ini dilakukan beberapa tahapan yaitu data *preprocessing*, uji asumsi, pengelompokan, dan interpretasi *cluster*. Data *preprocessing* yang dilakukan terdiri dari tiga yaitu transformasi data, deteksi *outlier*, dan normalisasi data. Periode waktu saat ini yang digunakan dalam transformasi data yaitu 1 Januari 2022. Transformasi data merubah dataset awal sebanyak jumlah 41.073 *row* data dengan empat variabel yaitu kontak id, id transaksi, *purchased date*, dan total *amount* menjadi 5.108 *row* data dengan variabel yang terbentuk yaitu kontak id, *length*, *recency*, *frequency*, dan *monetary*. *Script* yang digunakan untuk transformasi data menggunakan *tools* MySQL sebagai berikut:

```
SELECT
    `KONTAK ID`,
    timestampdiff(DAY, min(`PURCHASED DATE`), max(`PURCHASED DATE `)) as
    LENGTH,
    timestampdiff(DAY, min(`PURCHASED DATE`), '2022-01-01')) as RECENCY,
    COUNT(DISTINCT IDTRANSAKSI) as FREQUENCY,
    SUM(`TOTAL AMOUNT`) as MONETARY
FROM dataskripsi.datamentah2
GROUP BY
`KONTAK ID`
```

Pendeteksian *outlier* dalam penelitian ini menggunakan variabel *multivariate* yaitu menggunakan metode pengukuran jarak kuadrat mahanalobis. Perhitungan jarak kuadrat

mahanalobis setiap objek menggunakan Persamaan 1. Penelitian ini menggunakan empat variabel sehingga $p = 4$ dengan nilai $\alpha = 5\%$ sehingga objek dikatakan *outlier* jika jarak kuadrak mahanalobis objek lebih besar dari $\chi^2_{4(1-0.05)} = 9{,}488$. Hasil deteksi *outlier* menggunakan variabel *length*, *recency*, *frequency*, dan *monetary* dengan bantuan aplikasi GUI R didapatkan hasil bahwa 124 objek merupakan data *outlier*. Penelitian ini tetap mempertahankan data *outlier* dalam analisis berikutnya menggunakan metode *k-medoids*. *K-Medoids* merupakan metode *clustering* yang *robust* terhadap *outlier*.

Data memiliki satuan yang berbeda agar dalam pemodelan tidak menghasilkan hasil yang bias maka dilakukan normalisasi data. Normalisasi data pada variabel *recency* memiliki perhitungan yang berbeda yaitu perhitungan kembali dengan perhitungan $\left(1 - \hat{X}_{\text{recency}}\right)$. Normalisasi data tanpa adanya perhitungan kembali pada variabel *recency* menyebabkan perhitungan CLV mengalami *miss leading* dikarenakan semakin kecil nilai *recency* maka semakin baik berbanding terbalik dari tiga variabel lainnya.

Analisis *cluster* memiliki uji asumsi yang perlu dipenuhi. Asumsi analisis *cluster* yaitu sampel mewakili populasi dan multikolinearitas.

1. Sampel Mewakili Populasi
   Hasil pengujian KMO diperoleh nilai KMO sebesar 0,779. Nilai KMO yang didapat lebih besar dari 0,50 sehingga dapat disimpulkan bahwa sampel yang digunakan mewakili populasi atau asumsi sampel *representative* terpenuhi.
2. Non Multikolinearitas
   Hasil pengujian VIF diperoleh hasil bahwa semua parameter memiliki nilai VIF lebih kecil dari 10 yaitu variabel *length* memiliki VIF sebesar 4,003, *recency* memiliki VIF sebesar 2,178, *frequency* memiliki VIF sebesar 3,113, dan *monetary* memiliki VIF sebesar 1,329. Oleh karena itu, data tidak terindikasi adanya multikolinearitas sehingga asumsi non multikolinearitas terpenuhi.

Hasil dua asumsi analisis *cluster* telah terpenuhi sehingga data dapat dianalisis berikutnya menggunakan metode *k-medoids*. Metode *k-medoids* memiliki kelemahan salah satunya tidak dapat menentukan *cluster* optimal. *Cluster* optimal dapat dicari dengan menggunakan uji validasi *silhouette index*. Hasil perbandingan rata-rata nilai koefisien *silhouette* untuk mendapatkan *cluster* optimal dapat dilihat pada Tabel 1.

Tabel 1. Hasil Perhitungan Rata-rata Nilai Koefisien *Silhouette*

k	Ukuran Jarak	
	Euclidean	*Manhattan*
2	0,60	0,62
3	0,55	0,51
4	0,54	0,53
5	0,48	0,49
6	0,46	0,45
7	0,42	0,39
8	0,41	0,38
9	0,39	0,39
10	0,40	0,37

Hasil *cluster* menggunakan variabel model LRFM dengan metode *k-medoids* dibantu validasi *silhouette index* diperoleh bahwa *cluster* optimal yang terbentuk yaitu dua *cluster* dengan menggunakan ukuran jarak *manhattan* yang dapat dilihat pada Tabel 1. Hal tersebut dikarenakan memiliki rata-rata nilai koefisien *silhouette* tertinggi dibandingkan dengan kombinasi kriteria *cluster* yang lainnya.

Langkah selanjutnya yaitu melakukan pengelompokan pelanggan menggunakan metode *k-medoids* berdasarkan variabel model LRFM menggunakan *cluster* optimal yang didapat. Hasil pengelompokan yang didapatkan kemudian dikalikan dengan bobot setiap parameter model W-LRFM untuk menghitung CLV. Bobot setiap parameter pada setiap perusahaan memiliki pembobot yang berbeda berdasarkan hasil survei dan hasil analisis pembobotan menggunakan *Analytics Hierarchical Process* (AHP). Hasil survei dan hasil analisis pembobotan pada salah satu perusahaan farmasi di Jakarta didapat bahwa nilai W_L, W_R, W_F, dan W_M yaitu 0,16, 0,29, 0,47, dan 0,08. Hasil perhitungan CLV dapat dilihat pada Tabel 2.

Tabel 2. Hasil Perhitungan *Customer Lifetime Value* (CLV)

Cluster	W_L**Length*	W_R**Recency*	W_F**Frequency*	W_M**Monetary*	CLV
1	0,016	0,130	0,012	0,000	0,158
2	0,123	0,263	0,107	0,001	0,494

Tabel 2 merupakan tabel perhitungan CLV. Tabel tersebut menenjukan bahwa CLV *cluster* kedua memiliki CLV lebih tinggi dibandingkan *cluster* pertama sehingga *cluster* kedua menjadi prioritas utama dalam strategi pemasaran salah satu *brand* pada salah satu perusahaan farmasi di Jakarta. Langkah terakhir yaitu melakukan interpretasi dan profilisasi *cluster*. Interpretasi hasil *cluster* menggunakan nilai-nilai *centroid* variabel setiap *cluster* yang dibandingkan dengan rata-rata secara keseluruhan setiap variabel tersebut. *Centroid* adalah rata-rata nilai objek yang terdapat dalam *cluster* pada setiap variabel. Nilai *centroid* setiap variabel dapat dilihat pada Tabel 3.

Tabel 3. Nilai *Centroid* Setiap *Cluster*

Cluster	Pelanggan	*Length*	*Recency*	*Frequency*	*Monetary*
1	2.444	37	203	2	1.048.773
2	2.664	288	36	10	7.102.826
Rata-rata Keseluruhan		168	115	6	4.206.172

Pemetaan karakter pelanggan berdasarkan matriks kesetiaan pelanggan dilakukan dengan melihat Tabel 3. Tanda ↑ menandakan nilai *centroid* variabel pada *cluster* di atas nilai rata-rata keseluruh pelanggan pada variabel tersebut sedangkan tanda ↓ menandakan nilai *centroid* variabel pada *cluster* di bawah nilai rata-rata keseluruh pelanggan pada variabel tersebut. Hasil pemetaan karakter pelanggan setiap *cluster* dapat dilihat pada Tabel 4.

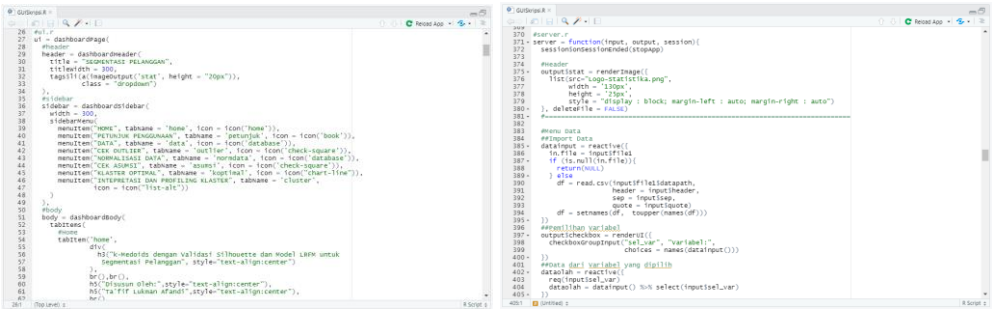
Tabel 4. Pemetaan Karakter Pelanggan

Cluster	Tipe LRFM	Karakteristik Pelanggan
1	$L\downarrow R\uparrow F\downarrow M\downarrow$	Pelanggan Tidak Berpotensi
2	$L\uparrow R\downarrow F\uparrow M\uparrow$	Pelanggan Setia

Tabel 3 dan 4 merupakan tabel hasil segmentasi pelanggan dua *cluster*. Hasil segmentasi pelanggan menyatakan bahwa pelanggan c*luster* 1 yang beranggota sebanyak 2.444 yang dikategorikan sebagai pelanggan yang tidak menyakinkan dan berpotensi menghilang (pelanggan tidak berpotensi) sedangkan pelanggan *cluster* 2 beranggota sebanyak 2.664 yang dikategorikan sebagai pelanggan dengan kesetiaan tinggi (pelanggan setia).

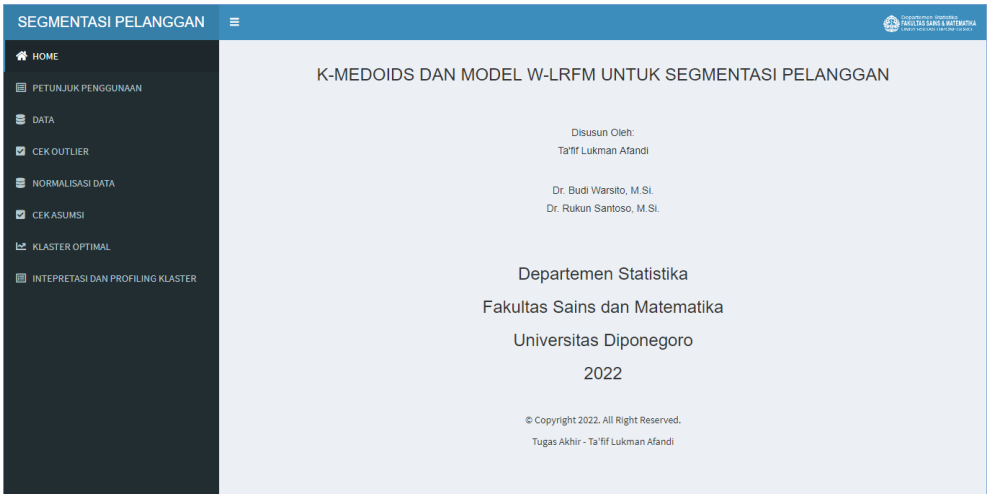
Algoritma *k-medoids* memiliki kompleksitas yang tinggi sehingga penelitian ini dilengkapi aplikasi *Graphical User Interface* (GUI) R. Pembuatan aplikasi GUI R digunakan untuk melakukan pengelompokan pelanggan menggunakan metode *k-kedoids* dan model W-LRFM (*Weighted-Length Recency Frequency Monetary*) dengan cara memasukan data,

memasukan nilai *input* numerik, dan menekan tombol. Pembuatan aplikasi GUI R diperlukan *package* R salah satunya yaitu *Shiny*. Komponen dalam pembuatan aplikasi GUI R terdiri dari dua komponen utama yaitu *User Interface* (UI) dan *Server*. Tampilan komponen *User Interface* dan *Server* dapat dilihat pada Gambar 3.



Gambar 2. Tampilan Program UI dan *Server*

Langkah berikutnya yaitu menjalankan aplikasi dengan cara mengklik pada tombol *Run App* pada sisi kanan atas sehingga muncul tampilan seperti pada Gambar 5.



Gambar 3. Tampilan Awal GUI R

5. KESIMPULAN

Hasil pengelompokan pelanggan salah satu *brand* perusahaan farmasi di Jakarta periode Januari 2021 – Desember 2021 menggunakan metode *k-medoids* dengan uji validasi *silhouette index* didapatkan *cluster* optimal yaitu menggunakan ukuran jarak *manhattan* dengan jumlah k *cluster* sebanyak dua *cluster*. Karakteristik hasil segmentasi dua *cluster* berdasarkan implementasi *k-medoids* dan model W-LRFM didapatkan bahwa *cluster* 1 memiliki anggota sebanyak 2.444 (47,8%) pelanggan dengan karakteristik pelanggan yang tidak berpotensial ($L \downarrow R \uparrow F \downarrow M \downarrow$) dan CLV sebesar 0,158 sedangkan *cluster* 2 memiliki anggota sebanyak 2.664 (52,2%) pelanggan dengan karakteristik pelanggan yang setia ($L \uparrow R \downarrow F \uparrow M \uparrow$) dan CLV sebesar 0,494. Pembuatan *Graphical User Interface* (GUI) R digunakan sebagai alat bantu dalam menganalisis pada penelitian pengelompokan pelanggan menggunakan *k-medoids* dan model W-LRFM dengan cara memasukan data, memasukan nilai *input* numerik, dan menekan tombol.

DAFTAR PUSTAKA

Arora, P., Deepali, dan Varshney, S. 2015. Analysis of K-Means and K-Medoids Algorithm for Big Data. *Procedia Computer Science*, Vol. 78, hh. 507-512. (https://www.sciencedirect.com/science/article/pii/S1877050916000971)

Artun, O. dan Levin, D. 2015. *Predictive Marketing Easy Ways Every Marketer Can Use Customer Analytics and Big Data*. United States of America. Penerbit Willey.

Brownlee, J. 2020. *Data Preparation for Machine Learning*. Machine Learning Mastery.

Buttle, F. dan Maklan, S. 2015. *Customer Relantionship Management, Third edition*. New York. Penerbit Butterworth-Heinemann.

Carneiro, R. dan Migueis, V. 2021. Applying Data Mining Techniques and Analytics Hierarchy Process to the Food Industry: Estimating Customer Lifetime Value. *Proceedings of the International Conference on Industrial Engineering and Operation Management* Vol. 125, hh. 266 – 277. (http://www.ieomsociety.org/brazil2020/papers/125.pdf)

Hair, J. F., Black W. C., Babin, B. J., dan Anderson, R. E. 2010. *Multivariate Data Analysis, Seventh edition*. New Jersey. Penerbit Prentice Hall International, Inc.

Johnson, R. A. dan Wichern, D. W. 2007. *Applied Multivariate Statisticak Analysis*. New Jersey. Penerbit Pearson.

Kandeil, D. A., Amani, A. S., dan Youssef, S. M. 2014. A Two-phase Clustering Analysis for B2B Customer Segmentation. *International Conference on Intelligent Networking and Collaborative Systems*, IEEE INcoS 2014, hh. 221-228. (https://sci-hub.se/10.1109/incos.2014.49).

Kaufman, L. dan Rousseeuw, P. J. 1990. *Finding Groups in Data an Introduction to Cluster Analysis*. New Jersey. Penerbit John Wiley & Sons Inc Publication.

Nugroho, S. 2008. *Statistika Multivariat Terapan*. Bengkulu. Penerbit UNIB Press.

Primartha, R. 2021. *Algoritma Machine Learning*. Bandung. Penerbit Informatika.

Santosa, B. dan Umam, A. 2018. *Data Mining dan Big Data Analytics*. Yogyakarta. Penerbit Media Pustaka.

Suyanto. 2017. *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung. Penerbit Informatika.

Tirta, I. M. 2014. Pengembangan E-Modul Statistika Terintegrasi dan Dinamik dengan R-shiny dan MathJax. *Universitas* Jember, hh. 223-232. (https://core.ac.uk/display/296257630).

Wei, J.T., Lin S.Y., Weng C.C., dan Wu H.H. 2012. A Case Study of Applying LRFM Model in Market Segmentation of a Children's Dental Clinic. *Elsevier*, Vol. 39, hh. 5529-5532. (https://www.sciencedirect.com/science/article/abs/pii/S0957417411016125).