

Penerapan Metode K-Means Dengan Metode Elbow Untuk Segmentasi Pelanggan Menggunakan Model RFM (Recency, Frequency & Monetary)

Adnan Burhan Hidayat Kiat^{*1}, Yufis Azhar², Vinna Rahmayanti³

^{1,2,3}Teknik Informatika/Universitas Muhammadiyah Malang

hidayatkiat123@gmail.com^{*1}, yufis.az@gmail.com², vinna.nastiti@gmail.com³

Abstrak

Segmentasi pelanggan pada perusahaan merupakan tindakan yang dapat mempermudah perusahaan dalam mengambil keputusan ke depan. Pada penelitian ini data yang digunakan berasal dari perusahaan otomotif, PT Hasjrat Abadi Ambon. Data yang dipakai terdiri dari data transaksi dan pelanggan kendaraan bermotor. Penerapan model RFM dapat mengelompokkan pelanggan-pelanggan berdasarkan nilai variabel Recency, Frequency dan Monetary. Hasil dari model RFM akan memperoleh status baru pada tiap pelanggan dari skala terbaik sampai terburuk. Pelanggan yang telah memiliki status akan dikelompokkan menggunakan metode K-Means menjadi beberapa Cluster(kelompok). Dalam menentukan jumlah Cluster yang optimal maka diterapkan metode Elbow. Algoritma yang digunakan dalam pembentukan Cluster terdiri dari Euclidean Distance dan Manhattan Distance. Kedua algoritma akan dibandingkan kualitas pembentukan Clusternya menggunakan metode Silhouette Coefficient. Hasil yang diberikan pada penelitian ini berupa data yang terbagi atas 5 kelompok dengan dilakukannya lima kali pengujian untuk menentukan centroid yang unggul. Cluster yang unggul akan dibuatkan visualisasi datanya untuk memudahkan perusahaan dalam mengambil keputusan. Berdasarkan penerapan Silhouette Coefficient, algoritma yang lebih unggul yaitu Manhattan Distance dengan nilai $s(i)$ sebesar 0.152695.

Kata Kunci: K-Means, Recency Frequency Monetary (RFM), Silhouette Coefficient, Segmentasi Pelanggan

Abstract

Customer segmentation at the company is an action that can facilitate the company in making decisions going forward. In this study the data used came from an automotive company, PT Hasjrat Abadi Ambon. The data used consists of transaction data and motor vehicle customers. The application of the RFM model can classify customers based on the value of the Recency, Frequency and Monetary variables. The results of the RFM model will obtain a new status on each customer from the best to the worst scale. Customers who already have status will be grouped using the K-Means method into several Clusters (groups). In determining the optimal number of Clusters, the Elbow method is applied. The algorithm used in Cluster formation consists of Euclidean Distance and Manhattan Distance. The two algorithms will be compared the quality of the Cluster formation using the Silhouette Coefficient method. The results given in this study are in the form of data divided into 5 groups by conducting five tests to determine superior centroids. Excellent clusters will be made of data visualization to facilitate the company in making decisions. Based on the application of Silhouette Coefficient, a superior algorithm is Manhattan Distance with value $s(i)$: 0.152695.

Keywords: K-Means, Monetary Frequency Recency (RFM), Silhouette Coefficient, Customer Segmentation

1. Pendahuluan

Pendekatan perusahaan dengan pelanggan merupakan hal penting untuk mempertahankan keuntungan perusahaan. Pemahaman atas perbedaan dari setiap pelanggan dibutuhkan untuk memahami kebutuhan dari tiap pelanggan berdasarkan pada data pelanggannya. Pemanfaatan dari data pelanggan dapat memberikan patokan kepada perusahaan untuk menentukan penjualan selanjutnya. Bisnis penjualan kendaraan sepeda motor, membutuhkan solusi untuk dibuatkan segmentasi pelanggan dalam rangka untuk

mengetahui perbedaan dari tiap pelanggan saat bertransaksi, contohnya pelanggan yang sering melakukan pembelian pada kurun waktu hari/bulan/tahun, atau pelanggan yang memberikan profit yang tinggi pada perusahaan. Segmentasi dari pelanggan tersebut dapat diolah informasinya dengan cara penerapan data *mining* [1].

Kurangnya pemahaman akan informasi penting pada data yang dimiliki merupakan salah satu kendala yang terjadi pada perusahaan PT Hasjrat Abadi. Data yang terdiri dari angka dan huruf pada data Pelanggan dan Transaksi mengandung informasi yang berguna untuk kepentingan perusahaan dalam manajemen penjualan produk kepada pelanggan. Kemudahan dalam memahami kondisi segmentasi pelanggan dapat diperoleh dengan menerapkan data *mining* pada rekaman data yang dimiliki. Pemanfaatan dari data *mining* dapat memberikan informasi yang berguna ke depan dan dapat dengan mudah dipahami dengan mengimplementasikan visualisasi data. Grafik yang terdiri dari pengelompokan data pelanggan secara mudah dapat dipahami bagaimana segmentasi pelanggan berdasarkan parameternya [2].

Dalam menjadikan data pelanggan menjadi suatu informasi yang berguna, maka dibutuhkan analisa *mining* pada kumpulan data yang telah terekap. PT. Hasjrat Abadi telah mengumpulkan data transaksi dan pelanggan sejak tahun 2012 sampai dengan 2019 sebanyak 6.490 data. Data tersebut akan dilakukan tahapan *preprocessing* untuk mengambil variabel tertentu yang dapat digunakan sebagai pembagian/segmentasi pelanggan. Setelah dilakukan tahapan *preprocessing* maka data akan di *mapping* terlebih dahulu menggunakan variabel yang ada pada model *RFM* (*Recency, Frequency, Monetary*). Pemodelan *RFM* sering digunakan untuk mengelompokkan waktu terakhir kunjungan, frekuensi dari kunjungan, dan pendapatan yang diperoleh pada perusahaan [3][4]. *RFM* memiliki kinerja bagaimana tiap pengelompokan akan diberikan skor (5-1), nilai skor tersebut akan digunakan untuk dianalisa menggunakan metode *K-Means* yang sering digunakan untuk *Clustering*(pengelompokan) data.

K-Means tergolong algoritma yang sering digunakan dalam mengelompokkan data. Nilai *K* (*Cluster*) akan ditentukan sebelum dibentuk *Cluster* nya. Penggunaan Metode *Elbow* pada beberapa penelitian diyakini baik untuk penentuan nilai *K* pada metode *K-Means* [5][6]. Metode *Elbow* memiliki kinerja dengan cara menguji coba satu per satu nilai *K* yang akan disimpan pada *Sum Square Error* (*SSE*), setelah ditemukan nilai *K Cluster* yang optimal maka *K Centeroid* akan menghasilkan titik pusat secara acak, dan tiap data akan dikalkulasi sampai membentuk suatu *Cluster* sesuai nilai *K Cluster* yang dibuat. Pada penulisan ini nilai yang akan di *Clustering* berasal dari hasil *mapping* pada model *RFM* [6]. Pada penelitian [1] [2] [4] [6] telah menjelaskan keuntungan dari penerapan Metode *K-Means*, Metode *K-Medoids* dengan model *RFM* memiliki kinerja yang sesuai dalam pembuatan segmentasi pelanggan berdasarkan dari data pelanggan. Alasan penulis tidak menggunakan metode *K-Medoids* dikarenakan pada hasil penelitian [7] dan penjelasan dasar dari metode *K-Medoids* [8] disimpulkan bahwa metode tersebut tidak unggul dalam menganalisa *Big-Data* atau sekumpulan data yang memiliki banyak variabel.

Pada penelitian ini penulis bertujuan untuk menerapkan Metode *K-Means* dengan penggabungan Metode *Elbow* dalam menentukan nilai *K* yang optimal. Dikarenakan hasil dari *K-Means* sangat berpengaruh terhadap *centroid* yang dipilih, maka analisa akan dilakukan sebanyak 5 kali pengujian dengan nilai *centroid* yang berbeda. Pembentukan segmentasi pelanggan akan menggunakan *dataset* dari hasil penerapan model *RFM*. Model *RFM* akan memberikan nilai skor pada setiap pelanggan berdasarkan nilai *Recency, Frequency, dan Monetary*. Skor yang telah diberikan dari penerapan model *RFM* akan disegmentasikan menggunakan metode *K-Means* dengan kedua algoritma perhitungan yaitu, *Euclidean Distance* dan *Manhattan Distance*. Kedua algoritma tersebut akan dibandingkan kualitas *clusternya* menggunakan metode *Silhouette Coefficient*. Pelanggan yang telah disegmentasikan akan diperlakukan masing-masing tindakan oleh pihak perusahaan, pelanggan dengan skor *RFM* yang rendah akan diberikan tindakan sehingga dapat berkunjung kembali ke perusahaan dan bertransaksi. Pelanggan yang skor menengah atau tinggi akan di lakukan tindakan yang lain juga kepada mereka untuk dapat mempertahankan tingkat loyalitas dari pelanggan. Hasil akhir yang diperoleh berupa informasi penting dari segmentasi pelanggan berdasarkan penerapan model *RFM* dan berbagai metode yang dipilih. Data segmentasi pelanggan akan dibentuk sebuah visualisasi data untuk memberikan keputusan kepada perusahaan dalam memberikan tindakan lanjut ke tiap-tiap pelanggan berdasarkan statusnya.

2. Metode Penelitian

2.1 Pengumpulan Data

Data yang dipakai pada penelitian ini merupakan data dari kantor PT Hasjrat Abadi Ambon. Perusahaan ini beroperasi dalam *Dealer* kendaraan bermotor dan mobil, data yang dipakai berupa data transaksi dan data pelanggan kendaraan bermotor. Data yang tersusun oleh perusahaan beroperasi secara harian, dan data yang terkumpul berupa hasil dari gabungan tiap-tiap cabang yang ada di Provinsi Maluku. Dari data yang terkumpul akan dibuatkan Clustering/pengelompokan data menjadi beberapa kelompok, pengelompokan tersebut bertujuan untuk memudahkan perusahaan dalam memahami data. Tujuan dari penggunaan data ini yaitu dalam menerapkan Segmentasi pelanggan yang akan dibentuk menggunakan keempat metode yaitu *K-Means*, *Silhouette Coefficient*, *Elbow*, dan *RFM*.

2.1.1 Sampel Data

Sampel data berisi dari data yang akan digunakan dalam penelitian ini, data yang dipakai berupa data transaksi dan data pelanggan. Kedua data akan digabungkan menjadi satu sumber data untuk memudahkan tahapan penelitian selanjutnya. Gabungan kedua data bertujuan dalam fitur tambahan untuk membuat visualisasi data, seperti data dari umur, gender, alamat, jenis motor, dll.

2.2 Preprocessing Data

Sebelum dilakukan tahapan analisa pada *dataset*, maka terlebih dahulu dilakukan tahapan data preprocessing, hal ini bertujuan untuk mengurangi error dalam tahapan analisa. *Dataset* terkadang memiliki kendala pada data yang hilang ataupun kekeliruan pada pengisian data. Nilai yang berada pada tiap variabel dalam kumpulan data dapat memiliki kekeliruan jika tidak ada format pengisian data. Penginputan data yang tidak konsisten seperti contohnya, pengisian tanggal, dapat berupa 01 Desember 2019 dan 12-01-2019, dan masih banyak lagi kemungkinan penginputan yang tidak sesuai pada tiap nilai di berbagai variabel data. Penerapan data preprocessing dapat membersihkan data yang kosong bila mana data tersebut tidak terlalu diperlukan dalam analisa ataupun data yang kosong hanya terdiri dari sedikit data. Data yang memiliki Noise pada kekeliruan penginputan atau data yang kosong dapat dibersihkan dengan menghapus data tersebut. Setelah data telah dibersihkan maka akan mempermudah tahapan berikutnya untuk proses analisa [9].

2.3 Selecting Data

Data yang dipakai pada pembentukan segmentasi pelanggan berupa data dari variabel Nama Pelanggan, Tanggal DO-Invoice, dan Harga SPK. Nama pelanggan digunakan dalam mengidentifikasi data berdasarkan nama pelanggan dan dipakai untuk membentuk ID pelanggan. Tanggal DO-Invoice memiliki fungsi dalam menentukan nilai *Recency* dan *requery* pada penerapan metode *RFM*. Harga SPK memiliki peran dalam penentuan nilai *Monetary* dari model *RFM* [1]. Data yang dipilih akan dioperasikan menggunakan python. Data akan dipilih berdasarkan header data atau melalui Indeks dari tabel data, dan dibuatkan variabel baru untuk menyimpan data yang telah dipilih.

2.4 K-Means

K-Means merupakan metode Clustering yang memiliki kinerja dalam mengelompokkan dasar data item dalam bentuk fitur yang baru pada grup K. Nilai K pada metode ini merupakan nilai ketentuan kelas yang akan terbentuk berdasarkan pada nilai yang di masukan pada nilai K itu sendiri. Penginputan nilai selalu dalam bentuk Integer yang positif. Pengelompokan pada metode *K-Means* akan dilakukan dengan cara mengurangi jumlah total kuadrat jarak antara Cluster centroid dan data [8]. Algoritma yang di terapkan pada *K-Means* seperti berikut [1].

- a. Menentukan Nilai K yang akan dibuat.
- b. Menentukan Nilai K-Centroid secara acak.
- c. Kalkulasi jarak pada tiap data ke tiap Centroid.

Kalkulasi tersebut dapat menggunakan jarak *Euclidean* dan *Manhattan* dengan Persamaan 1 dan Persamaan 2 sebagai berikut.

1. Euclidean Distance

$$dist_{xy} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (1)$$

2. Manhattan Distance

$$dist_{xy} = |x_{ik} - x_{jk}| \quad (2)$$

- d. Data akan di kelompokkan berdasarkan jarak terdekat antara data dan centroid.
 e. Pengulangan proses c dan d sampai keseluruhan objek tidak dapat diklasifikasikan lagi.

Dalam penelitian ini penulis berencana dalam menentukan nilai K dengan menggunakan Metode *Elbow* untuk mencari nilai yang optimal dan membandingkan perhitungan jarak antara *Euclidean Distance* dan *Manhattan Distance*. Perbandingan tersebut akan di lihat dari sebaik apa kedua algoritma ini mengelompokkan data untuk tiap cluster dan akan di ambil nilai mean pada tiap cluster untuk dibandingkan kedua algoritma. Perbandingan ini akan menggunakan metode *Silhouette Coefficient* untuk mencari kualitas tiap cluster berdasarkan nilai $s(i)$ nya.

2.5 Elbow

Metode *Elbow* merupakan metode pendukung yang dapat menentukan nilai K yang optimal pada penerapan metode *K-Means*. Penerapan metode ini memiliki fokus pada persentase varian sebagai fungsi dari jumlah Cluster. Untuk mencari nilai K yang optimal maka nilai K akan dicek satu persatu dan akan dicatat nilai *SSE (Sum Square Error)*. Algoritma dalam mendapatkan nilai *SSE* yaitu seperti Persamaan 3 berikut.

$$SSE = \sum_{k=1}^k \sum_{xi \in Sk} \|Xi - Ck\|^2 \quad (3)$$

Nilai *SSE* adalah jumlah rata-rata dari Jarak *Euclidean* pada setiap titik terhadap Centroid. Pada diagram bila nilai turun secara signifikan dan membuat lekukan dari garis pada titik awal dan setelahnya maka nilai dari K telah di temukan [5].

2.6 Recency Frequency Monetary (RFM)

Merupakan Model dari penyusunan segmentasi pelanggan yang terdiri dari tiga variabel dalam tolak ukurnya. Variabel pada model *RFM* terdiri dari *Recency*, *Frequency* dan *Monetary*. *Recency* merupakan variabel penilaian dari rentan waktu kedatangan seorang pelanggan dari kedatangan terakhir sampai waktu sekarang. Data yang akan digunakan sebagai variabel *Recency* berupa data tanggal Do-Invoice yang merupakan rekapan dari kedatangan pelanggan dalam menyetujui pembelian. *Frequency* merupakan variabel penilaian dari jumlah kedatangan pelanggan ke perusahaan. Data yang akan dipakai untuk variabel *Frequency* sama seperti variabel *Recency* yaitu tanggal do-invoice. *Monetary* merupakan variabel penilaian dari total transaksi yang diberikan pelanggan pada perusahaan. Pada variabel ini data yang akan digunakan merupakan data harga SK, yaitu harga dari total unit pembelanjaan. Penerapan model *RFM* ini bertujuan dalam membentuk segmentasi pelanggan yang pada akhirnya akan dimanfaatkan dalam memberikan informasi yang tepat untuk penentuan treat kepada tiap pelanggan. Perusahaan melakukan tindakan ini demi meningkatkan tingkat kelayakitan dari pelanggannya [3][10]. Pada penelitian ini penulis berencana dalam menerapkan metode *K-Means* bersamaan dengan metode *Elbow* sebagai pengoptimalan nilai Cluster yang akan diterapkan juga dengan model *RFM* dalam penyusunan segmentasi pelanggan pada perusahaan PT Hasjrat Abadi Ambon, data yang dipakai berupa data transaksi pelanggan dari tahun 2012 sampai 2019.

2.7 Silhouette Coefficient

Merupakan metode yang memiliki kinerja dalam mencari nilai pada kumpulan data yang telah tersusun pada suatu Cluster. Data yang dicarikan merupakan nilai $s(i)$, yaitu nilai dari

pengukuran kualitas suatu data yang tersusun pada Cluster yang dibuat. Sebelum mencapai nilai $s(i)$ maka terlebih dahulu mencari nilai $a(i)$ yang merupakan nilai rata-rata jarak dari data i dengan data lain yang ada pada Cluster yang sama, juga nilai $d(i,C)$ yang merupakan nilai rata-rata jarak data i dengan keseluruhan Cluster atau tiap-tiap Cluster yang lain, setelah nilai $d(i,C)$ ditemukan maka akan dicari nilai minimum dari hasil tiap-tiap rata-rata jarak pada Cluster lain, nilai tersebut merupakan nilai dari $b(i)$. Rumus-rumus berikut merupakan Persamaan 4, Persamaan 5, Persamaan 6, dan Persamaan 7 dari *Silhouette Coefficient*.

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \quad (4)$$

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \quad (5)$$

$$b(i) = \min_{C \neq A} d(i, C) \quad (6)$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (7)$$

Pada penelitian ini penulis berencana dalam menggunakan metode *Silhouette Coefficient* dalam menentukan nilai data pada tiap Cluster untuk membantu penulis dalam memastikan Cluster yang dibentuk telah memenuhi kualitasnya. Dikarenakan pengelompokan menggunakan metode *K-Means* dan algoritma dari Euclidean *Distance* juga *Manhattan Distance* memerlukan pemilihan Centroid yang berperan sebagai titik nilai acuan pada pengelompokan Cluster. Pada dasarnya nilai centroid dipilih secara random/acak, maka dengan menggunakan metode *Silhouette Coefficient* ini tiap Cluster yang dibentuk akan dicari nilai $s(i)$ nya untuk menentukan kualitas dari Cluster tersebut. Penulis juga berencana dalam menentukan algoritma mana yang memiliki pengelompokan Cluster yang lebih unggul antara *Euclidean Distance* dengan *Manhattan Distance*. Dengan hasil nilai $s(i)$ yang telah ditemukan pada tiap-tiap Cluster akan dicari nilai rata-rata dari keseluruhan Clusternya dan akan dibandingkan kualitas dari kedua algoritma tersebut yang akan diacuhkan dengan skala dari +1 sampai -1. Bila hasil rata-rata nilai $s(i)$ mendekati angka 1 maka Cluster yang dibentuk tergolong baik atau berkualitas, begitu pula sebaliknya bila nilai $s(i)$ mendekati -1 [11][12].

3. Hasil Penelitian dan Pembahasan

Data akan dianalisa menggunakan IDE Jupiter Notebook Python dengan metode *K-Means* yang terbagi atas dua algoritma perhitungan jarak, yaitu *Euclidean Distance* dan *Manhattan Distance*. Sebelum data dikelompokkan menggunakan metode *K-Means*, maka terlebih dahulu dilakukan pemberian status pelanggan menggunakan model *RFM* dan menentukan jumlah kluster yang optimal menggunakan metode *Elbow*. Data yang telah tersusun akan dibuatkan visualisasinya untuk mempermudah mencerna informasi yang diperoleh.

3.1. Hasil Penerapan Model *RFM*

Untuk memperoleh nilai *Recency* maka nilai tertinggi dari data tanggal / hari terakhir terjadinya pendataan transaksi, akan dijadikan sebagai nilai yang akan dikurangi dengan tanggal kedatangan terakhir pelanggan (nilai max kedatangan pelanggan). Nilai *Frequency* didapatkan dengan dihitung jumlah pelanggan melakukan transaksi dan untuk nilai *Monetary* didapatkan dengan ditotalkan keseluruhan transaksi masing-masing pelanggan, ditunjukkan pada Tabel 1.

Tabel 1. Pemberian Nilai *RFM* untuk Tiap Pelanggan

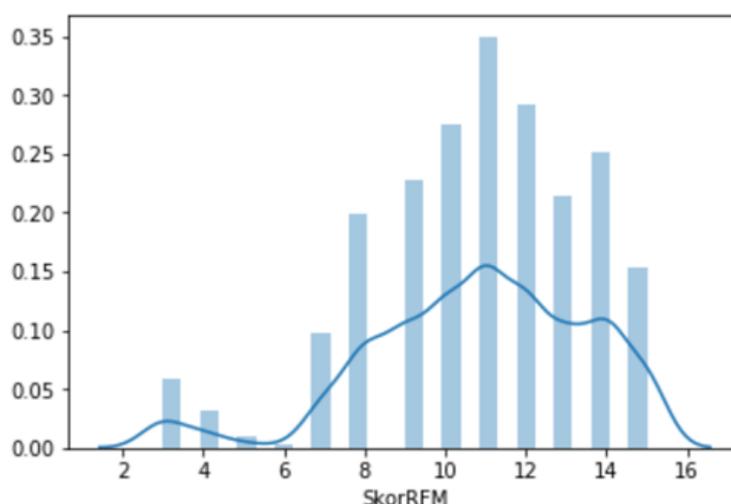
No.	Pelanggan	<i>Recency</i>	<i>Frequency</i>	<i>Monetary</i>
1	Costumer1	632	1	20950000
2	Costumer2	946	1	18550000
3	Costumer3	307	1	65400000
4	Costumer4	155	1	22850000
5	Costumer5	1017	1	18850000
6	Costumer6	1117	1	18350000

7	Costumer7	1149	1	20800000
8	Costumer8	723	1	30550000
9	Costumer9	1250	1	21000000
10	Costumer10	910	1	18650000
...

Hasil yang diperoleh membentuk data baru, data ini akan digunakan untuk menentukan Skor yang akan diberikan pada tiap-tiap pelanggan. Skor yang akan diberikan berupa urutan dari angka 1 sampai 5, bila skor yang didapati yaitu angka 5 maka tergolong kurang baik dan bila nilai mendekati skor 1 maka nilai tersebut tergolong bagus. Untuk menentukan skor pada tiap pelanggan, maka digunakan skala sebagai patokannya. Tiap data diberikan skala nilai dari nilai 1 sampai 0 dengan pembagian nilai jarak antara kedua nilai sebanyak 6 kali, maka skala yang diperoleh berupa angka 1, 0.8, 0.6, 0.39, 0.19, dan 0. Nilai skala yang terbagi atas 6 bagian akan diterapkan pada nilai variabel *Recency*, *Frequency*, dan *Monetary*. Dengan Skor yang diperoleh akan ditotalkan untuk mendapat SkorRFM, total skor akan digunakan untuk memberikan status pelanggan yang terdiri dari 5 kategori, yaitu *Titanium*, *Diamond*, *Gold*, *Silver* dan *Bronze*. Seperti ditunjukkan Tabel 2.

Tabel 2. Pemberian Skor RFM dan Status untuk Tiap Pelanggan

No.	Pelanggan	Recency	Frequency	Monetary	R	F	M	SkorRFM	Status
1	Costumer1	632	1	20950000	2	5	3	10	Diamond
2	Costumer2	946	1	18550000	3	5	5	13	Silver
3	Costumer3	307	1	65400000	2	5	1	8	Titanium
4	Costumer4	155	1	22850000	1	5	2	8	Titanium
5	Costumer5	1017	1	18850000	4	5	5	14	Bronze
6	Costumer6	1117	1	18350000	4	5	5	14	Bronze
7	Costumer7	1149	1	20800000	3	5	3	11	Gold
8	Costumer8	723	1	30550000	4	5	1	10	Diamond
9	Costumer9	1250	1	21000000	3	5	3	11	Gold
10	Costumer10	910	1	18650000	2	5	5	12	Gold
...



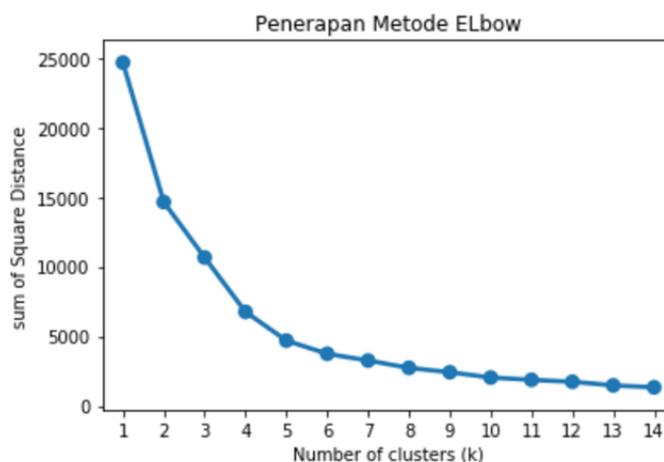
Gambar 1. Plot SkorRFM pada Keseluruhan Pelanggan

Berdasarkan Gambar 1, hasil penerapan model *RFM* pada *dataset*, diketahui bahwa status terbanyak dari keseluruhan pelanggan adalah SkorRFM bernilai 11 dengan status *Gold*. Dikarenakan segmentasi pelanggan ditentukan dari statusnya dan status tersebut hanya berdasarkan nilai SkorRFM, maka penulis berencana dalam membentuk Cluster/kelompok baru dari tiap pelanggan yang menggunakan *dataset* dari nilai R, F, dan M sebagai data yang akan di kelompokkan berdasarkan algoritma perhitungan jarak tiap-tiap datanya untuk memperoleh segmentasi pelanggan yang lebih akurat. Dalam membuat Cluster baru maka akan diterapkan

metode *K-Means* dengan kedua algoritma perhitungan jarak, yaitu *Euclidean Distance* dan *Manhattan Distance*.

3.2 Hasil Penerapan Metode *Elbow*

Pada Gambar 2, dalam mencari nilai optimal menggunakan metode *Elbow* maka, tiap data akan dicarikan nilai *SSE (Sum Square Error)* pada perbandingan Cluster = 1 sampai seterusnya. Hasil dari *SSE* tiap data akan divisualisasikan untuk menentukan Cluster manakah yang memiliki nilai optimal.



Gambar 2. Hasil Nilai SSE tiap Cluster

Dari hasil penerapan metode *Elbow* informasi yang diterima yaitu *Cluster* dengan nilai yang optimal berada pada *Cluster* 5, yang merupakan titik dengan kemiringan nilai *SSE* sempurna. Maka dalam penerapan metode *Elbow* dapat disimpulkan bahwa nilai yang optimal untuk menerapkan jumlah *Cluster* yaitu sebanyak 5.

3.3 Hasil Penerapan Metode *K-Means*

Setelah nilai optimal *Cluster* ditemukan maka dalam penerapan metode *K-Means* ini akan menggunakan jumlah *Cluster* sebanyak 5 bagian. Algoritma yang akan digunakan dalam mengelompokkan data-data yaitu, *Euclidean Distance* dan *Manhattan Distance*. Kedua algoritma tersebut akan dibandingkan kualitas dari pengelompokan *Cluster* yang telah dibuat menggunakan metode *Silhouette Coefficient*.

3.3.1 *Euclidean Distance*

Dari hasil penerapan metode *K-Means* dengan algoritma *Euclidean Distance* sebanyak lima kali pengujian, maka hasil yang diperoleh Tabel 3 sebagai berikut.

Tabel 3. Penerapan Metode *K-Means* dengan Algoritma *Euclidean Distance*

No.	Pelanggan	GrupRFM	SkorRFM	Status	Uji1	Uji2	Uji3	Uji4	Uji5
1	Costumer1	235	10	<i>Diamond</i>	2	1	3	2	1
2	Costumer2	355	13	<i>Silver</i>	4	4	3	3	1
3	Costumer3	351	8	<i>Titanium</i>	3	1	2	1	5
4	Costumer4	152	8	<i>Titanium</i>	3	1	2	1	5
5	Costumer5	455	14	<i>Bronze</i>	4	2	1	3	4
6	Costumer6	455	14	<i>Bronze</i>	4	2	1	3	4
7	Costumer7	353	11	<i>Gold</i>	2	3	3	2	1
8	Costumer8	451	10	<i>Diamond</i>	1	3	4	4	2
9	Costumer9	353	11	<i>Gold</i>	2	3	3	2	1
10	Costumer10	255	12	<i>Gold</i>	2	4	3	2	1
...

Pengujian yang dilakukan sebanyak lima kali ini akan digunakan sebagai patokan pada metode *Silhouette Coefficient* untuk dianalisa centroid yang menghasilkan kelompok data yang

berkualitas. Berdasarkan hasil analisa sebelumnya, maka dapat diringkas untuk mempermudah dalam memahami isi informasinya. Berikut merupakan tabel dari keseluruhan 5180 data yang telah dibagikan mencari 5 Cluster pada hasil pengujian ke 3, seperti pada Tabel 4.

Tabel 4. Ringkasan Hasil Analisa Pengujian ke 3

Cluster.	Pelanggan	R	F	M	SkorRFM	Titanium	Diamond	Gold	Silver	Bronze
1	1212	5	5	5	13-15	0	0	0	242	970
2	1105	1	5	2	7-9	1105	0	0	0	0
3	1492	2	5	4	10-13	0	416	847	202	0
4	1117	4	5	2	9-13	144	244	661	68	0
5	254	2	1	1	3-7	254	0	0	0	0

Cluster dengan status pelanggan terbaik berada pada Cluster 5. Pada Cluster 5 memiliki nilai rata-rata *RFM* yang terkecil(baik) dan memiliki skor *RFM* 3 sampai 7. Total pelanggan yang berada pada Cluster 2 secara rata-rata memiliki status *Titanium*, walaupun Cluster 2 memiliki pelanggan yang lebih besar pada status *Titanium*, akan tetapi nilai skor *RFM* pada Cluster 2 hanya terdiri dari pelanggan dengan skor 7 sampai 9. Dapat disimpulkan bahwa Cluster yang memiliki pelanggan dengan loyalitas tertinggi berada pada Cluster 2 dan yang terendah pada Cluster 1 dengan status *Silver* dan *Bronze*.

3.3.2 Manhattan Distance

Dari hasil penerapan metode *K-Means* dengan algoritma *Manhattan Distance* sebanyak lima kali pengujian, maka hasil yang diperoleh Tabel 5 dan Tabel 6 sebagai berikut.

Tabel 5. Penerapan Metode *K-Means* dengan Algoritma *Manhattan Distance*

No.	Pelanggan	GrupRFM	SkorRFM	Status	Uji1	Uji2	Uji3	Uji4	Uji5
1	Costumer1	253	10	<i>Diamond</i>	3	3	3	3	4
2	Costumer2	355	13	<i>Silver</i>	5	5	2	3	3
3	Costumer3	351	8	<i>Titanium</i>	2	2	1	2	4
4	Costumer4	152	8	<i>Titanium</i>	2	2	1	2	4
5	Costumer5	455	14	<i>Bronze</i>	5	5	2	5	2
6	Costumer6	455	14	<i>Bronze</i>	5	5	2	5	2
7	Costumer7	353	11	<i>Gold</i>	4	3	3	3	3
8	Costumer8	451	10	<i>Diamond</i>	4	4	4	4	4
9	Costumer9	353	11	<i>Gold</i>	4	3	3	3	3
10	Costumer10	255	12	<i>Gold</i>	3	3	3	3	3
...

Tabel 6. Ringkasan Hasil Analisa Pengujian ke 3

Cluster.	Pelanggan	R	F	M	SkorRFM	Titanium	Diamond	Gold	Silver	Bronze
1	1086	2	5	2	7-9	1086	0	0	0	0
2	1414	4	5	5	13-15	0	0	0	444	970
3	1453	2	5	4	9-12	163	416	847	0	0
4	973	4	5	2	10-13	0	244	661	68	0
5	254	2	1	1	3-7	254	0	0	0	0

Hasil dari penerapan algoritma *Manhattan Distance* menunjukkan bahwa Cluster yang dapat dikategorikan unggul dari nilai *RFM* nya yaitu pada Cluster 5 dengan pelanggan yang berstatus *Titanium* sebanyak 254 orang yang memiliki skor *RFM* berkisaran dari angka 3 – 7. Sama halnya dengan hasil analisa *Euclidean Distance*, pada Cluster ke 1 dapat dilihat bahwa jumlah pelanggan dengan status *Titanium* mencapai 1086 orang. Cluster 1 yang memiliki jumlah pelanggan yang banyak dengan status yang baik belum tentu lebih unggul dengan Cluster 5 karna perbedaan pada nilai *RFM* dan skor *RFM* yang tidak sebaik Cluster 5. Cluster yang dikategorikan sebagai tingkat loyalitas rendah berada pada Cluster 2, yang mendominasi dari pelanggan berstatus *Silver* dan *Bronze*.

3.4 Hasil Penerapan Metode *Silhouette Coefficient*

Dalam mencari hasil Cluster yang terbaik pada penerapan algoritma *Euclidean* dan *Manhattan Distance*, maka dilakukan tahapan analisa secara berulang untuk menentukan hasil Cluster yang terbaik. Dalam penelitian ini tahapan untuk mencari Cluster yang terbaik dilakukan lima kali pengujian yang berbeda pada penentuan centroid untuk membentuk Cluster. Setelah dilakukan pengujian dengan lima centroid yang berbeda maka hasil dari analisa tersebut akan dicarikan salah satu yang terbaik dalam membentuk Cluster. Dalam mencari hasil yang terbaik maka diterapkan metode *Silhouette Coefficient*. Metode ini dapat memberikan hasil nilai $s(i)$ pada tiap Cluster yang dibentuk, nilai $s(i)$ merupakan penentu pada suatu Cluster yang dibuat dengan centroid yang telah ditentukan memiliki hasil yang baik atau sebaliknya. Nilai yang dihasilkan menggunakan metode *Silhouette Coefficient* bila mendekati angka 1 maka dapat dikatakan pembentukan Cluster tersebut berkualitas dan bila mendekati angka -1 maka hasil dari pengelompokan tergolong kurang baik. Dengan mengambil nilai rata-rata pada tiap Cluster dari hasil pengujian maka dapat digunakan sebagai patokan hasil pengujian mana yang memiliki kualitas Cluster terbaik. Tabel 7 dan Tabel 8 berikut merupakan centroid yang dibentuk pada 5 kali pengujian dengan algoritma *Euclidean* dan *Manhattan*.

Tabel 7. Lima Pengujian Centroid dengan Algoritma *Euclidean Distance*

Cluster.	C1	C2	C3	C4	C5
Uji 1	[4. 5. 2.]	[2. 5. 4.]	[1. 5. 2.]	[4. 5. 5.]	[2.1.1.]
Uji 2	[2. 5. 2.]	[5. 5. 5.]	[4. 5. 2.]	[2. 5. 4.]	[2.1.1.]
Uji 3	[5. 5. 5.]	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 2.]	[2.1.1.]
Uji 4	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 5.]	[4. 5. 2.]	[2.1.1.]
Uji 5	[2. 5. 4.]	[4. 5. 2.]	[2. 1. 1.]	[5. 5. 5.]	[1.5.2.]

Tabel 8. Lima Pengujian Centroid dengan Algoritma *Manhattan Distance*

Cluster.	C1	C2	C3	C4	C5
Uji 1	[1, 1, 1]	[2, 5, 2]	[3, 5, 3]	[4, 5, 4]	[5, 5, 5]
Uji 2	[1, 1, 1]	[2, 5, 2]	[3, 5, 3]	[4, 5, 3]	[5, 5, 5]
Uji 3	[2, 5, 2]	[5, 5, 5]	[3, 5, 3]	[4, 5, 3]	[1, 1, 1]
Uji 4	[1, 5, 2]	[2, 5, 3]	[3, 5, 4]	[4, 5, 4]	[5, 5, 5]
Uji 5	[5, 5, 5]	[4, 5, 4]	[3, 5, 3]	[2, 5, 2]	[1, 1, 1]

Data yang telah dianalisa pada kedua algoritma dengan pengujian sebanyak lima kali akan digunakan untuk mencari kualitas centroid pada tiap Cluster dari masing-masing algoritma. Data satu-persatu akan dianalisa dengan patokan dari tiap-tiap hasil uji sebagai Clusternya untuk mencari nilai variabel yang ada dalam metode ini. Tabel 9, Tabel 10, dan Tabel 11 berikut adalah hasil dari penerapan metode *Silhouette Coefficient*.

Tabel 9. Hasil Mencari Nilai $s(i)$ dari Kelima Pengujian pada Algoritma *Manhattan Distance*.

Cluster.	C1	C2	C3	C4	C5	Nilai $s(i)$
Uji 1	[4. 5. 2.]	[2. 5. 4.]	[1. 5. 2.]	[4. 5. 5.]	[2.1.1.]	0,149693
Uji 2	[2. 5. 2.]	[5. 5. 5.]	[4. 5. 2.]	[2. 5. 4.]	[2.1.1.]	0,152695
Uji 3	[5. 5. 5.]	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 2.]	[2.1.1.]	0,152695
Uji 4	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 5.]	[4. 5. 2.]	[2.1.1.]	0,137423
Uji 5	[2. 5. 4.]	[4. 5. 2.]	[2. 1. 1.]	[5. 5. 5.]	[1.5.2.]	-0,098291

Tabel 10. Hasil Mencari Nilai $s(i)$ dari Kelima Pengujian pada Algoritma *Euclidean Distance*.

Cluster.	C1	C2	C3	C4	C5	Nilai $s(i)$
Uji 1	[4. 5. 2.]	[2. 5. 4.]	[1. 5. 2.]	[4. 5. 5.]	[2. 1. 1.]	-0,011493
Uji 2	[2. 5. 2.]	[5. 5. 5.]	[4. 5. 2.]	[2. 5. 4.]	[2. 1. 1.]	-0,011898
Uji 3	[5. 5. 5.]	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 2.]	[2. 1. 1.]	-0,001540
Uji 4	[1. 5. 2.]	[2. 5. 4.]	[4. 5. 5.]	[4. 5. 2.]	[2. 1. 1.]	-0,011493
Uji 5	[2. 5. 4.]	[4. 5. 2.]	[2. 1. 1.]	[5. 5. 5.]	[1. 5. 2.]	-0,001540

Tabel 11. Perbandingan Hasil Nilai Silhouette Euclidean dan Manhattan Distance.

Algoritma	Hasil Uji	C1	C2	C3	C4	C5	Nilai s(i)
<i>Euclidean</i>	Uji 3	[5, 5, 5.]	[1, 5, 2.]	[2, 5, 4.]	[4, 5, 2.]	[2, 1, 1.]	-0.001540
<i>Manhattan</i>	Uji 3	[2, 5, 2]	[5, 5, 5]	[3, 5, 3]	[4, 5, 3]	[1, 1, 1]	0.152695

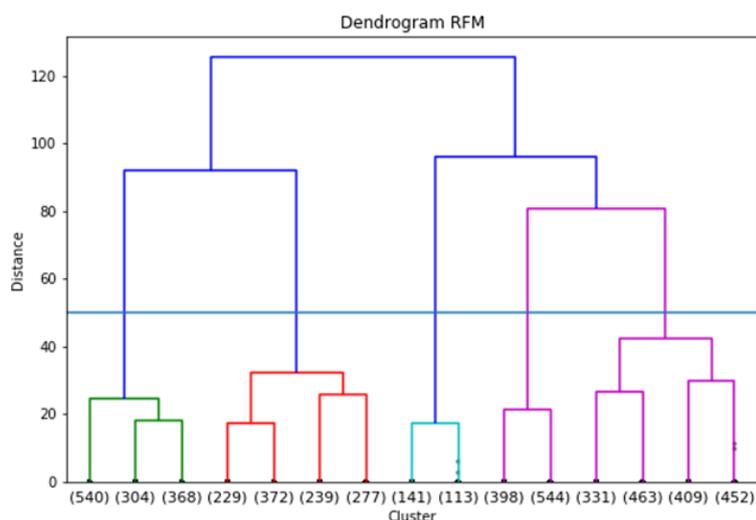
Dapat disimpulkan dari penerapan metode *Silhouette Coefficient* pada kelima pengujian nilai centroid yaitu, algoritma yang lebih unggul dalam mengelompokkan data-data *RFM* menjadi beberapa Cluster adalah *Manhattan Distance* dengan nilai sebesar 0.152695. Maka dalam pembuatan visualisasi akan menggunakan hasil dari Cluster uji yang ke 3 *Manhattan*.

Berdasarkan hasil analisa yang membuat algoritma *Manhattan* lebih unggul dalam pengelompokan data dikarenakan rumusan yang berbeda dari kedua algoritma. *Euclidean Distance* menghitung dengan rumusan akar kuadrat dari data, hal ini memiliki efek yang sangat signifikan bila data pada variabel memiliki *Outlier*, sedangkan *Manhattan Distance* menghitung dengan rumusan nilai absolut dari hasil pengurangan kedua data, hal ini menyebabkan *Outlier* yang berada pada data tidak akan berpengaruh banyak pada hasil akhir analisa. *Outlier* yang menyebabkan analisa *Euclidean* menjadi rendah berada pada variabel *F(Frequency)* pada dataset *RFM*. Penyebab variabel *F* memiliki *Outlier* dikarenakan rata-rata kedatangan pelanggan hanya sekali di perusahaan, hal ini menyebabkan model *RFM* memberikan skor $F = 5$ pada 75% pelanggan, dan pelanggan yang sering berkunjung ke perusahaan berkisar 25%.

3.5. Visualisasi Data

Data yang telah dianalisa akan dibentuk dalam visualisasi berupa *Pie Chart*, Diagram, *Histogram*, *Dendrogram*, dll. Hal ini diperlukan untuk mempermudah memahami isi informasi yang terkandung dalam hasil analisa. Pada pembentukan visualisasi, data yang digunakan merupakan hasil pengelompokan dari metode *K-Means* yang terdiri dari dua algoritma, *Euclidean Distance* dan *Manhattan Distance*. Hasil algoritma yang perlu difokuskan yaitu *Manhattan Distance*, berdasarkan hasil penelitian membuktikan bahwa, *Manhattan* tergolong algoritma yang lebih baik dalam pengelompokan data. Berikut adalah perbandingan dari kedua hasil analisa.

3.5.1. Dendrogram dari Hasil Clustering *Manhattan* dan *Euclidean*



Gambar 3. Dendrogram dari keseluruhan Dataset

Pengelompokan Gambar 3, dari tiap data terdiri dari berbagai varian nilainya, kesamaan pada tiap-tiap data akan membentuk kelompok baru dengan jarak yang terdekat. Pada dataset *RFM* terdiri dari tiga variabel yang berisi rentan nilai dari 1 sampai 5, pada masing-masing variabel memiliki varian yang berbeda pada jumlah kesamaan tiap datanya. Berdasarkan dendrogram di atas, varian data variabel *Recency* terdiri dari dua kelompok, yaitu kelompok berwarna hijau dan merah, variabel *Frequency* memiliki sedikit varian terhadap datanya yang berada pada kelompok berwarna Cyan atau biru langit, dan variabel *Monetary* terdiri dari 2 kelompok berwarna ungu.

Dengan ditentukan garis perpotongan pada jarak 50 maka varian dari *dataset* ini dapat membentuk cluster sebanyak lima kelompok seperti pada keterangan Gambar 4 dan Gambar 5, sama seperti hasil analisa yang diberikan dari metode *Elbow*.

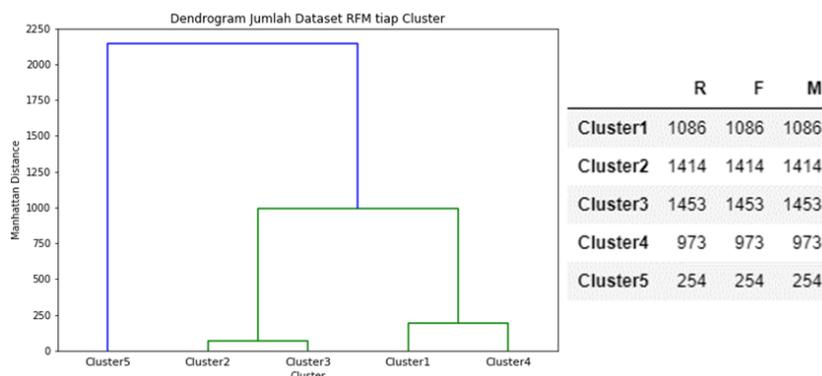


Gambar 4 Dendrogram Nilai Rata-Rata Cluster Manhattan

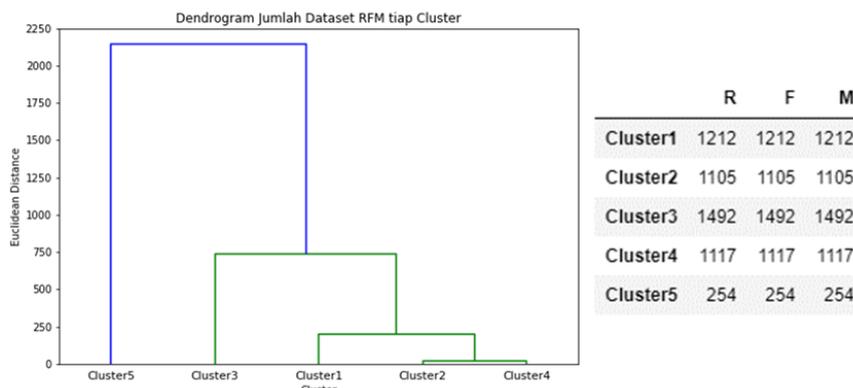


Gambar 5 Dendrogram Nilai Rata-Rata Cluster Euclidean

Perbandingan dari kedua algoritma tidak memiliki perbedaan yang signifikan. Nilai rata-rata *dataset* dari kedua algoritma memiliki hasil yang hampir sama pada tiap cluster, hal yang membedakan dari kedua algoritma ini hanya terdiri dari pemilihan centroid untuk tiap clusternya. Berdasarkan dendrogram jarak *Manhattan*, Cluster 1 dengan cluster 3 memiliki kemiripan nilai yang cukup dekat, begitu juga dengan cluster 2 dan 4, dengan demikian cluster-cluster tersebut dikelompokkan menjadi kelompok yang baru. Kedua kelompok ini juga memiliki kedekatan nilai yang serupa dengan jarak yang tidak begitu jauh. Cluster dengan kemiripan nilai yang jauh berada pada cluster 5 dengan nilai rata-ratanya yang terdiri dari skor 1 dibandingkan dengan cluster lainnya.



Gambar 6. Dendrogram Jumlah Dataset Cluster Manhattan

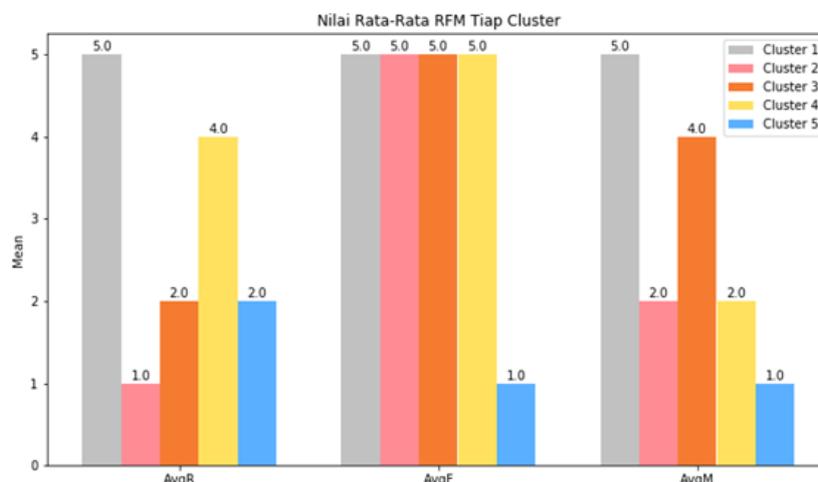


Gambar 7. Dendrogram Jumlah Dataset Cluster Euclidean

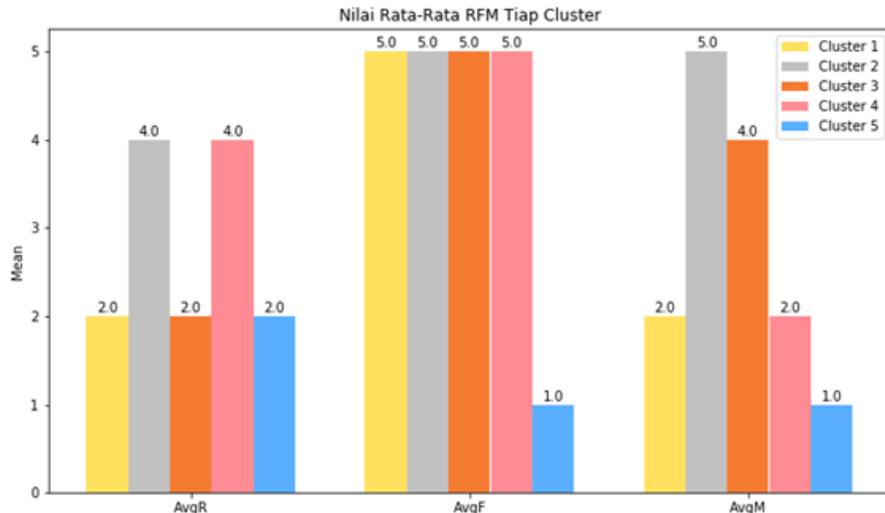
Pada Gambar 6 dan Gambar 7, jumlah *dataset* dari kedua algoritma memiliki perbedaan pada tiap-tiap clusternya. Dendrogram *Manhattan* menunjukkan bahwa cluster 2 dan cluster 3 memiliki kedekatan pada jumlah *dataset*nya, dengan perbandingan jarak sebesar 34, hal yang sama pada cluster 1 dan cluster 4 dengan perbandingan jarak sebesar 113. Kedua kelompok tersebut memiliki perbandingan jarak sebesar 404. Pada cluster 5 dengan jumlah *dataset* senilai 254, memiliki kedekatan nilai yang sangat jauh dengan perbandingan jarak sebesar 977.5 pada kedua cluster yang telah dikelompokkan. Berbeda dengan *Manhattan*, *Euclidean* memiliki varian data yang lebih pada tiap clusternya. Cluster 2 dan 4 memiliki kedekatan nilai yang sangat dekat dengan perbandingan jarak sebesar 12, begitu juga dengan cluster 1 yang memiliki kedekatan nilai dengan kedua -cluster tersebut dengan perbandingan jarak sebesar 101, dari kedua kelompok tersebut memiliki perbandingan jarak sebesar 330.5. Sama dengan *Manhattan*, cluster 5 memiliki jarak yang jauh dari gabungan cluster yang lain, dengan perbandingan jaraknya sebesar 1072.75.

3.5.2 Perbandingan Nilai Rata-Rata RFM Tiap Cluster

Dari perbandingan kedua hasil visualisasi Gambar 8 dan Gambar 9 ini dapat dilihat bahwa urutan Cluster terbaik sampai terendah berada pada posisi acak pada algoritma Euclidean Distance dikarenakan centroid yang dipilih secara acak melalui fungsi bawaan dan Manhattan dengan centroid yang di input secara acak. Warna yang digunakan pada tiap Cluster memberi gambaran status baru. Berawal dari warna biru yang menandakan kelompok Cluster dengan pelanggan berstatus Titanium, warna pink menandakan pelanggan berstatus Diamond, dan sisanya kuning, Silver dan orange memberikan status Gold, Silver dan Bronze. Nilai rata-rata dari keduanya memiliki sedikit perbedaan pada beberapa Cluster, seperti Cluster yang berstatus Diamond pada hasil analisa Euclidean yang memiliki nilai rata-rata R sebesar 1, juga pada Cluster lainnya yang memiliki perbedaan nilai rata-rata terkecuali pada Cluster dengan status Titanium.

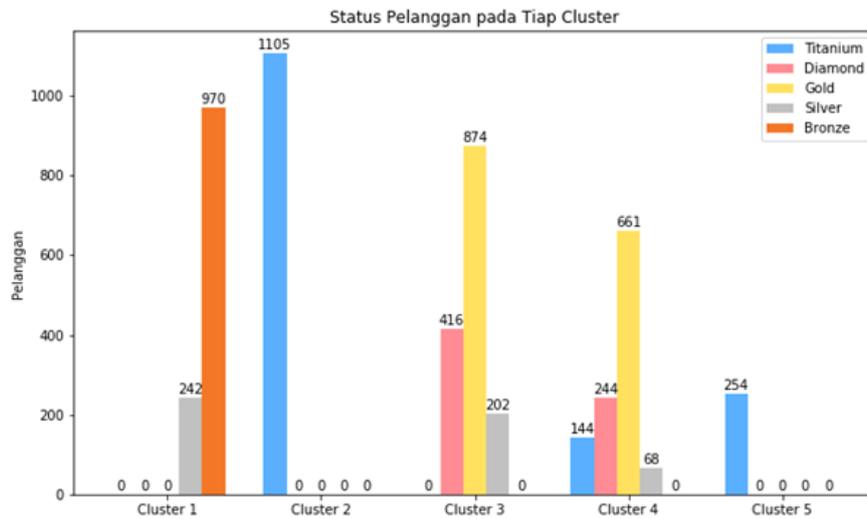


Gambar 8 Nilai rata-rata RFM hasil Euclidean Distance

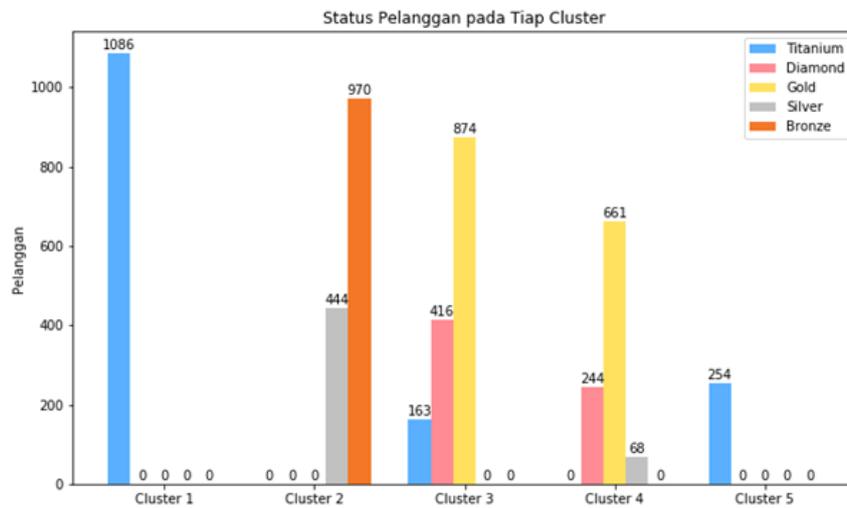


Gambar 9. Nilai rata-rata RFM hasil Manhattan Distance

3.5.3 Perbandingan Status Pelanggan pada Tiap Cluster



Gambar 10. Status Pelanggan hasil Euclidean Distance



Gambar 11. Status Pelanggan hasil Manhattan Distance

Kedua algoritma Gambar 10 dan Gambar 11 tersebut masih memiliki kesamaan pada pengelompokan pada tiap data-data, terutama pada Cluster 5 *Manhattan* dan Cluster 5 *Euclidean*. Dari kedua algoritma dapat dilihat bahwa Cluster 2 *Euclidean* dan Cluster 1 *Manhattan* memiliki banyak sekali pelanggan dengan status *Titanium*, akan tetapi Cluster yang lebih unggul ialah Cluster dengan pelanggan *Titanium* berjumlah 254 orang, hal ini dikarenakan skor *RFM* pada pelanggan tersebut berkisaran 3 – 7 yang berarti lebih unggul. Cluster yang lain dari kedua algoritma tidak terlalu signifikan menunjukkan perbedaannya, untuk perbedaan kecil berada pada Cluster 4 *Euclidean* dengan Cluster 4 *Manhattan* yang terlihat bahwa algoritma *Euclidean* mengelompokkan pelanggan hampir secara keseluruhan status, di mana terdiri dari *Titanium*, *Diamond*, *Gold* dan *Silver* dibandingkan dengan *Manhattan* yang hanya mengelompokkan *Titanium*, *Diamond* dan *Gold*.

3.5.4 Perbandingan Pengelompokan Cluster disertai Gender

Untuk kedua algoritma Gambar 12 dan Gambar 13 ini menghasilkan perbandingan yang tidak terlalu signifikan. Nilai rata-rata gender pada tiap Cluster yang lebih mendominasi berada pada pihak laki-laki. Sesuai dengan pie plot Cluster 5 *Euclidean* dan Cluster 5 *Manhattan*, menunjukkan bahwa pelanggan dengan tingkat loyalitas tertinggi yaitu sebesar 4.9% dari keseluruhan pelanggan dan pelanggan dengan loyalitas terendah berada pada Cluster 2 *Manhattan* sebanyak 27.3% pelanggan dan *Euclidean* berada pada Cluster 1 sebanyak 28.8% Pelanggan. Hal ini menunjukkan bahwa pelanggan di perusahaan PT Hasjrat Abadi Ambon memiliki kerentanan dalam kehilangan pelanggan yang berstatus loyalitas terendah. Dengan ini maka diperlukannya pendekatan dengan pelanggan berupa pemberian discount, give away dan tindakan lain yang dapat menarik kembali tingkat loyalitas pelanggan tersebut.

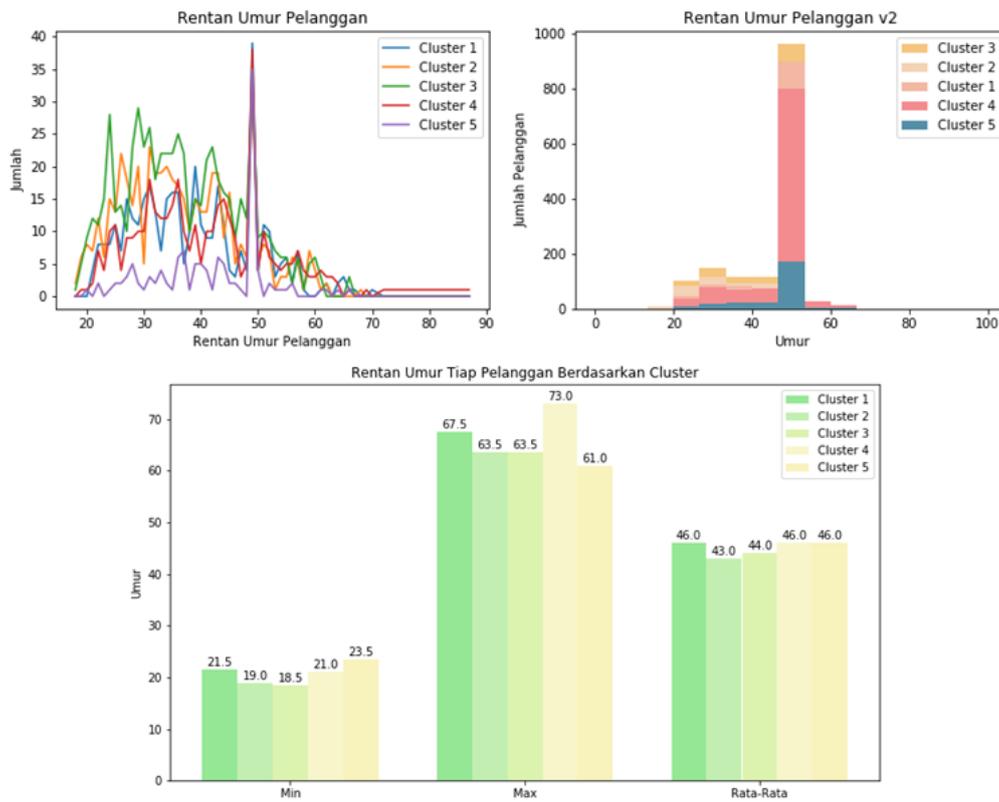


Gambar 12. Pengelompokan Cluster hasil Euclidean Distance

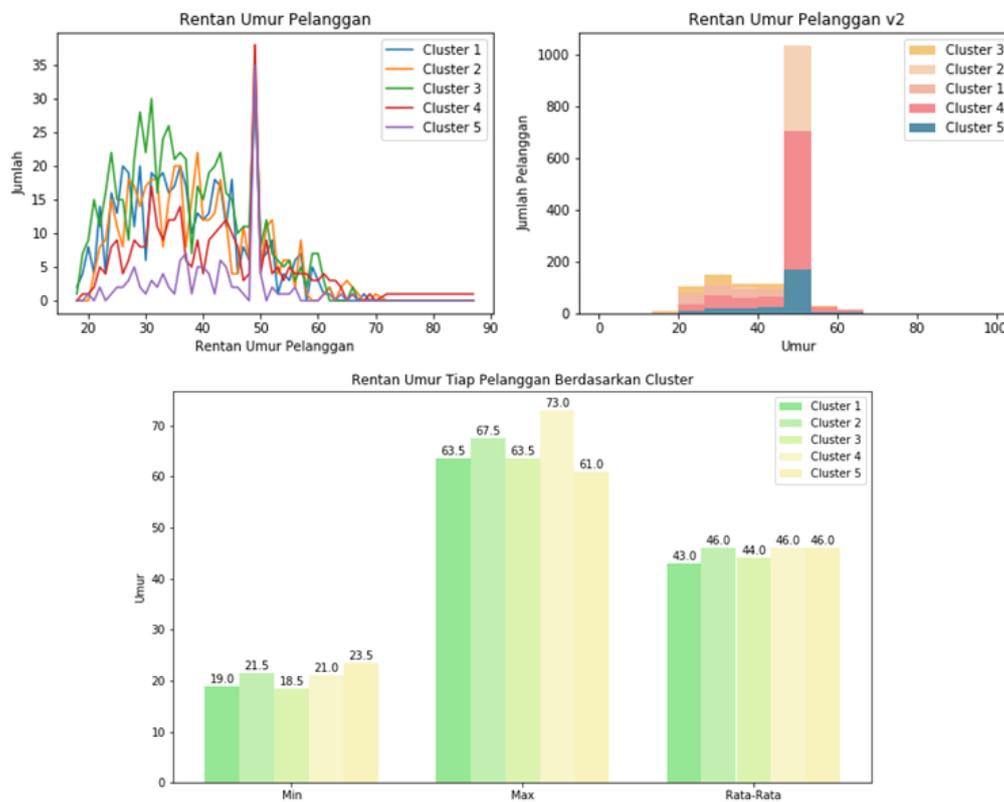


Gambar 13. Pengelompokan Cluster hasil Manhattan Distance

3.5.5. Perbandingan Rentan Umur Pelanggan pada Tiap Cluster



Gambar 14. Rentan Umur Pelanggan hasil Euclidean Distance

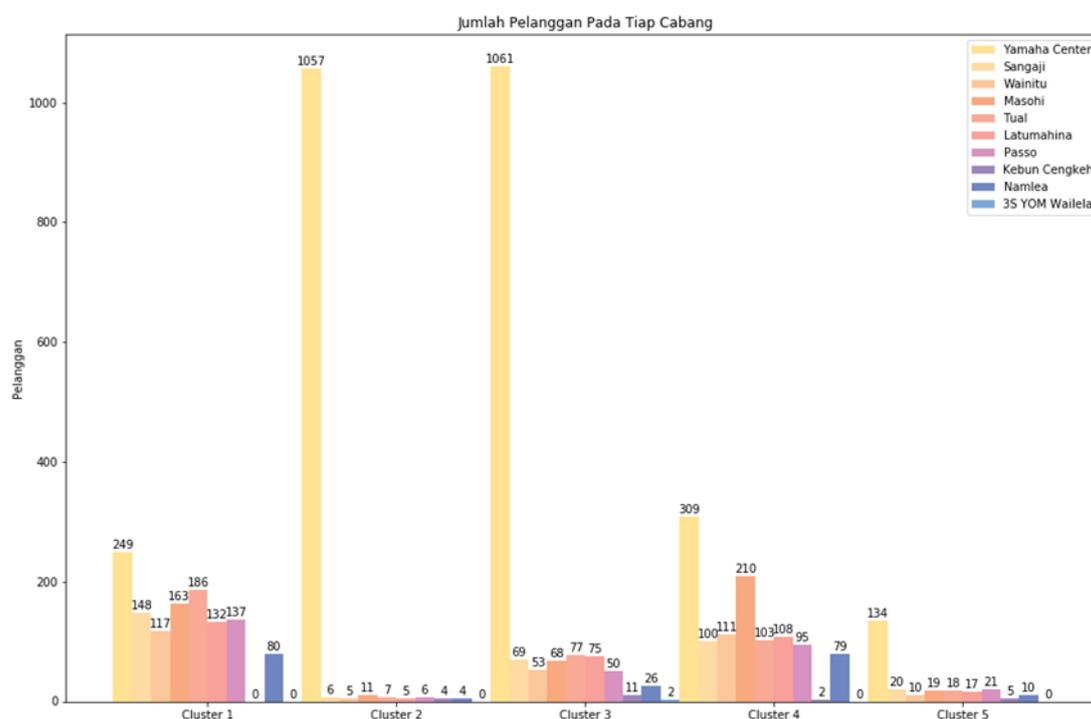


Gambar 15. Rentan Umur Pelanggan hasil Manhattan Distance

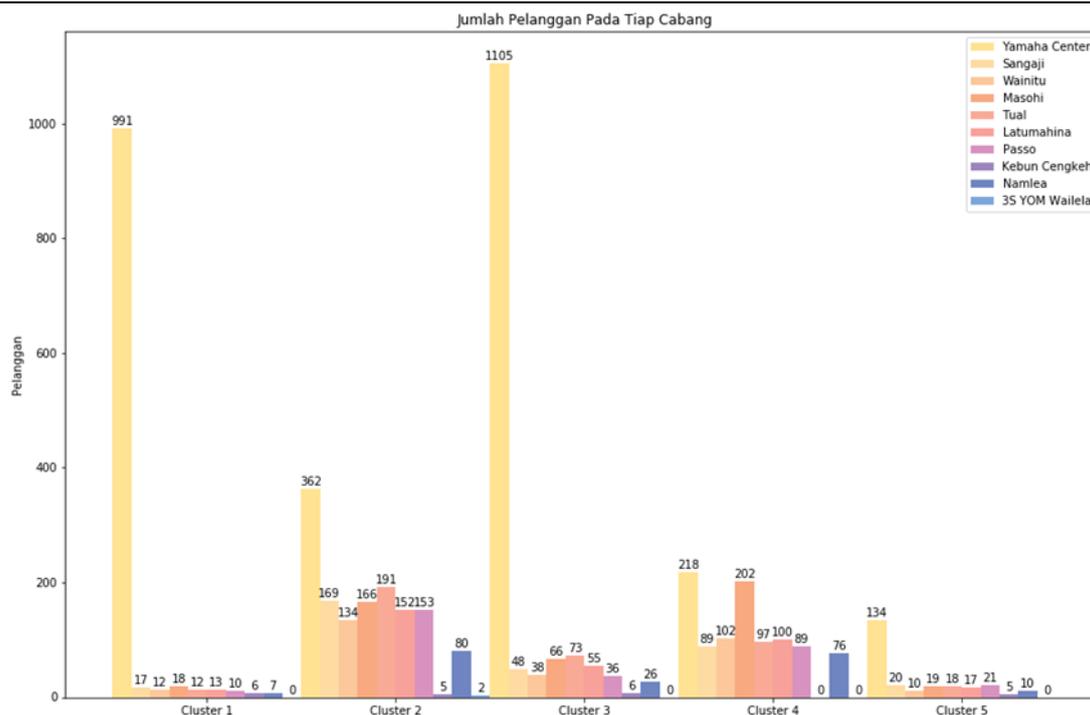
Dari kedua hasil analisa algoritma *Manhattan* dan *Euclidean* pada Gambar 14 dan Gambar 15, memberikan hasil yang tidak begitu berbeda pada keduanya. Pada visualisasi dari Rentan Umur Pelanggan 1 dan 2 dari kedua algoritma memberikan hasil yaitu, pelanggan terbanyak berada pada rentan umur 49 tahun. Dari keseluruhan Cluster memberikan hasil bahwa rentan umur yang dominan yaitu berumur 49 tahun, dan sisanya berkisaran dari 20 sampai 50 untuk tingkat tertinggi lainnya. Pada keseluruhan Cluster yang memiliki nilai minimal pada rentan umur pelanggan yaitu berada pada Cluster 3 *Manhattan* dengan umur berkisaran 18 tahun dan *Euclidean* berada pada Cluster 3, yang memiliki umur maksimal berada pada Cluster 4 dengan kisaran 73 tahun sama halnya dengan *Euclidean*. Nilai rata-rata umur pelanggan pada keseluruhan Cluster berkisaran 43 – 46 tahun, hal ini dapat disimpulkan bahwa pelanggan pada PT Hasjrat Abadi rata-ratanya berusia 43 – 46 tahun dan yang lebih banyak melakukan transaksi yaitu pelanggan dengan rentan usia 49 tahun.

3.5.6. Perbandingan Pelanggan pada tiap Cabang sesuai Cluster

Pada kedua visualisasi Gambar 16 dan Gambar 17 terlihat bahwa, cabang yang terdiri dari banyak pelanggan berada pada cabang Yamaha Center itu sendiri. Perbandingan dari pengelompokan kedua algoritma ini tidak terlalu banyak perbedaan. Pada hasil visualisasi *Manhattan* menandakan bahwa cabang dengan status Titanium terbesar berada pada Cluster ke 1, sedangkan untuk *Euclidean* Distance cabang dengan jumlah pelanggan terbanyak berada pada Cluster 2. Cabang-cabang yang memiliki tingkat loyalitas rendah berada pada cabang Kebun Cengkeh dan 3S YOM Wailela. Maka kesimpulan dari visualisasi tersebut yaitu, PT Hasjrat Abadi Ambon dapat mempertahankan sistem manajemen pada cabang di Yamaha Center, dan dapat memberikan solusi kepada cabang lain untuk meningkatkan jumlah kedatangan pelanggan. Cabang yang tergolong rentan pelanggan seperti cabang Kebun Cengkeh dan 3S YOM Wailela mungkin dapat diutamakan dalam peningkatan jumlah pelanggannya.



Gambar 16. Pelanggan pada tiap Cabang hasil *Euclidean* Distance



Gambar 17. Pelanggan pada tiap Cabang hasil Manhattan Distance.

4. Kesimpulan

Setelah berhasil melakukan tahapan analisa, implementasi dan visualisasi dari hasil penerapan metode *Recency*, *Frequency* dan *Monetary (RFM)* dan *K-Means*, memberikan hasil berupa data yang telah dikelompokkan berjumlah 5 kelompok dengan dilakukan pengujian sebanyak lima kali, masing-masing kelompok terdiri dari pelanggan-pelanggan yang memiliki status baru untuk mempermudah perusahaan dalam mengambil tindakan selanjutnya. Pelanggan dengan tingkat loyalitas tertinggi berada pada Cluster 5 dengan pelanggan berstatus *Titanium*, keseluruhan pelanggan tersebut berjumlah 254 pelanggan. Dengan diterapkan metode *Elbow* maka pembentukan Cluster pada metode *K-Means* jadi semakin mudah untuk memperoleh kelompok data yang berkualitas. Model *RFM* berhasil memberikan skor *RFM* kepada tiap pelanggan dan status baru untuk dibandingkan tingkat loyalitas mereka. Dalam menerapkan metode *K-Means*, algoritma *Euclidean Distance* diterapkan secara otomatis dengan library bawaan python dengan baik, dan algoritma *Manhattan Distance* dilakukan secara manual secara berulang untuk menemukan titik centroid yang lebih baik. Dengan dibandingkan kedua algoritma tersebut menggunakan metode *Silhouette Coefficient* maka algoritma *Manhattan* yang lebih unggul dibandingkan *Euclidean* dengan perhitungan rata-rata nilai $s(i)$ pada keseluruhan Cluster yang dibuat. Algoritma *Manhattan Distance* memperoleh skor $s(i)$ sebesar 0.152695, yang merupakan hasil dari nilai terbaik kelima pengujian begitu juga *Euclidean Distance* memperoleh skor $s(i)$ sebesar -0.001540. Maka dalam penelitian ini algoritma yang lebih unggul yaitu *Manhattan Distance*.

Berdasarkan hasil yang diperoleh dan telah disimpulkan terkait pembuatan segmentasi pelanggan PT Hasrat Abadi Ambon, menggunakan model *RFM*, dan metode *Elbow*, *K-Means* (*Euclidean Distance* & *Manhattan Distance*) dan *Silhouette Coefficient*, maka adapun beberapa saran seperti berikut:

1. Dapat dilakukan pengujian secara berkala dalam menentukan nilai Centroid menggunakan metode *Silhouette Coefficient* dengan memilih nilai centroid secara acak dan bervariasi untuk memperoleh nilai $s(i)$ yang terbaik.
2. Implementasi yang dilakukan dapat dikembangkan menjadi suatu program dalam memperoleh informasi data dengan mudah.

Daftar Notasi

x_{ik}	: Data ke-i.
x_{jk}	: Centroid ke-j.
X_j	: Data ke-i pada Cluster.
μ_i	: Nilai Mean Cluster.
$dist_{xy}$: Jarak antara Centroid dan Data.
$a(i)$: Nilai rata-rata jarak antara Data ke-i dengan Data yang lain dalam Cluster yang sama.
$d(i, C)$: Nilai rata-rata jarak antara Data ke-i dengan Data yang lain dalam Cluster ke-C.
$b(i)$: Nilai Terkecil dari Data ke-i pada tiap Cluster-C.
$s(i)$: Nilai <i>Silhouette Coefficient</i> .

Referensi

- [1] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, and E. A. Pratama, "Customer Segmentation based on *RFM* model and Clustering Techniques With *K-Means* Algorithm," *2018 Third Int. Conf. Informatics Comput.*, pp. 1–6.
- [2] T. Choi, H. K. Chan, S. Member, and X. Yue, "Recent Development in Big Data Analytics for Business Operations and Risk Management," pp. 1–12, 2016.
- [3] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "*RFM* ranking – An effective approach to customer segmentation," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018.
- [4] M. Namvar, "A Two Phase Clustering Method for Intelligent Customer Segmentation," 2010.
- [5] D. Marutho, S. H. Handaka, and E. Wijaya, "The Determination of Cluster Number at k-mean using *Elbow* Method and Purity Evaluation on Headline News," *2018 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 533–538, 2018.
- [6] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using *K-Means* Clustering," *2018 Int. Conf. Comput. Tech. Electron. Mech. Syst.*, pp. 135–139, 2018.
- [7] M. Aryuni and E. Miranda, "Customer Segmentation in XYZ Bank using *K-Means* and *K-Medoids* Clustering," *2018 Int. Conf. Inf. Manag. Technol.*, no. September, pp. 1–9, 2018.
- [8] K. M. A. Patel and P. Thakral, "The Best Clustering Algorithms in Data Mining," no. 2, pp. 2042–2046, 2016.
- [9] S. Mohanavalli and S. K. K. B, "Survey of Pre-processing Techniques for Mining Big Data," 2017.
- [10] M. Tavakoli, "Customer Segmentation and Strategy Development based on User Behavior Analysis , *RFM* model and Data Mining Techniques : A Case Study," no. April 2019, 2018.
- [11] M. Anggara, H. Sujiani, and N. Helfi, "Pemilihan *Distance* Measure Pada *K-Means* Clustering Untuk Pengelompokkan Member Di Alvaro Fitness," *J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1–6, 2016.
- [12] H. Wang *et al.*, "An approach for improving *K-Means* algorithm on market segmentation," *2010 Int. Conf. Syst. Sci. Eng. ICSSE 2010*, pp. 368–372, 2010.