BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka

2.1.1 Segmentasi Pelanggan

Segmentasi pelanggan memegang peranan penting dalam mengelola hubungan dengan pelanggan. Hal ini memungkinkan perusahaan untuk merancang dan menetapkan strategi yang berbeda untuk memaksimalkan nilai pelanggan. Segmentasi pelanggan mengacu pada pengelompokan pelanggan ke dalam kategori yang berbeda berdasarkan karakteristik bersama seperti usia, lokasi, kebiasaan belanja dan sebagainya. Demikian pula, pengelompokan berarti menyatukan sesuatu sedemikian rupa sehingga jenis hal yang serupa tetap berada dalam kelompok yang sama (Hung, P. D.,dkk, 2019).

Menurut Philip Kotler (2009), segmentasi pasar terdiri dari sekelompok pelanggan yang memiliki sekumpulan kebutuhan dan keinginan yang serupa. Segmentasi pasar merupakan suatu usaha untuk meningkatkan ketepatan pemasaran perusahaan. Segmen pasar terdiri dari kelompok besar yang dapat diidentifikasikan dalam sebuah pasar dengan keinginan, daya beli, lokal geografis, perilaku pembelian dan kebiasaan pembelian yang serupa.

Menurut Solomon (2011), segmentasi adalah proses membagi pasar yang lebih besar menjadi bagian-bagian yang lebih kecil berdasarkan satu atau lebih karakteristik bersama yang bermakna. Dengan melaksanakan segmentasi pelanggan, kegiatan pemasaran dapat dilakukan lebih terarah dan sumber daya yang dimiliki perusahaan dapat digunakan secara lebih efektif dan efisien dalam rangka memberikan kepuasan bagi pelanggan. Selain itu perusahaan dapat melakukan program-program pemasaran yang terpisah untuk memenuhi kebutuhan khas masing-masing segmen pelanggan.

Dalam bukunya yang berjudul "*Marketing Management*", Philip Kotler (2009), mengemukakan dasar membagi segmentasi menjadi empat variabel segmentasi utama bagi pelanggan, yaitu:

1) Segmentasi Geografis

Segmentasi geografis dilakukan dengan mengelompokkan pelanggan menjadi bagian pasar menurut skala wilayah atau letak geografis seperti: negara, provinsi, kota atau lingkungan. Segmentasi geografis ini penting mengingat kebutuhan maupun kegunaan suatu produk dan jasa selalu akan berbeda-beda tergantung pada lokasi, keadaan, maupun cuaca. Selain itu, segmentasi geografis digunakan untuk mengklasifikasikan pasar berdasarkan lokasi yang akan mempengaruhi biaya operasional dan jumlah permintaan secara berbeda.

2) Segmentasi Demografis

Dalam segmentasi demografis, pengelompokan konsumen berfokus pada variabel-variabel demografis seperti usia, ukuran keluarga, siklus kehidupan keluarga, jenis kelamin, penghasilan, pekerjaan, agama, ras, generasi, kewarganegaraan, dan kelas sosial. Variabel-variabel demografis adalah dasar yang paling populer untuk membedakan kelompok-kelompok pelanggan dan sering terkait dengan kebutuhan dan keinginan konsumen.

3) Segmentasi Psikografis

Segmentasi psikografis dilakukan dengan mengelompokkan konsumen atau pembeli menjadi bagian pasar berdasarkan sifat psikologis/kepribadian, pola atau gaya hidup (*life style*) atau nilai. Sebagai contoh, segmen pasar masyarakat yang bergaya hidup konsumtif dan mewah berbeda dengan segmen pasar masyarakat yang bergaya hidup produktif dan hemat yang mementingkan kualitas dengan harga yang relatif murah. Konsumen dalam kelompok demografi yang sama bisa memiliki profil psikografis yang sangat berbeda.

4) Segmentasi Perilaku

Dalam segmentasi perilaku, perusahaan membagi pembeli menjadi beberapa kelompok berdasarkan pengetahuan, sikap, penggunaan atau respon terhadap suatu produk.

Agar segmen pasar dapat bermanfaat maka harus memenuhi beberapa karakteristik:

a) Terukur (*Measurable*). Ukuran, daya beli, dan profil segmen harus dapat diukur meskipun ada beberapa variabel yang sulit diukur.
b) Substansial (*Substantial*). Segmen harus cukup besar dan menguntungkan untuk dilayani.
c) Dapat diakses (*Accessible*). Segmen harus dapat dijangkau dan dilayani secara efektif.
d) Dapat didiferensiasi (*Differentiable*). Segmen-segmen dapat dipisahkan secara konseptual dan memberikan tanggapan yang berbeda terhadap elemen-elemen dan bauran pemasaran yang berbeda.
e) Dapat ditindaklanjuti (*Actionable*). Program yang efektif dapat dibuat untuk menarik dan melayani segmen.

2.1.2 Model RFM

Model RFM merupakan salah satu metode yang dapat dilakukan untuk menganalisis perilaku pelanggan pada segmentasi pasar. Dengan mengukur perilaku pelanggan, model RFM dapat menganalisis dan memprediksi lebih lanjut perilaku pelanggan dalam suatu database (Wei, J. T., Lin, S. Y., Yang, Y. Z., & Wu, H. H, 2020). Melalui adopsi model RFM, seorang pengambil keputusan dapat secara efektif mengidentifikasi pelanggan yang berharga dan akan digunakan sebagai pengembangan strategi pemasaran yang efektif (Shihab, Afroge, & Mishu, 2019). Model RFM ini diperkenalkan pertama kali oleh Arthur Huges pada tahun 1994. Model RFM ini merupakan proses penilaian berdasarkan perilaku pelanggan yang dilihat dari waktu transaksi terakhir pelanggan (*recency*), jumlah transaksi (*frequency*), dan uang yang dikeluarkan (*monetary*) (Parikh, 2020; Monalisa, 2019). Keuntungan model RFM terletak pada relevansinya selama beroperasi pada beberapa variabel yang dapat diamati dan bersifat objektif. Variabel ini digolongkan menurut 3 kriteria:

1) *Recency* yaitu variabel untuk mengukur nilai pelanggan berdasarkan rentang waktu (tanggal, bulan, tahun) dari transaksi terakhir pelanggan selama periode

analisis. Semakin kecil rentang waktunya, semakin baik nilai *recency* (R).

2) *Frequency* yaitu variabel untuk mengukur nilai pelanggan berdasarkan jumlah transaksi yang dilakukan oleh pelanggan pada perusahaan selama periode yang analisis. Semakin besar jumlah transaksi yang dilakukan, semakin besar nilai *frequency* (F).

3) *Monetary* yaitu variabel untuk mengukur nilai pelanggan berdasarkan jumlah uang yang dikeluarkan atau dibelanjakan pelanggan untuk perusahaan selama periode analisis. Semakin besar jumlah uang yang dikeluarkan oleh pelanggan, semakin besar nilai *monetary* (M).

Berdasarkan nilai RFM yang digunakan dalam analisis segmentasi pelanggan, Tsiptsis & Chorianopoulos (2011) membagi karakteristik pelanggan menjadi 6 jenis, dapat dilihat pada Tabel 2.1.

Tabel 2.1 Karakteristik Pelanggan

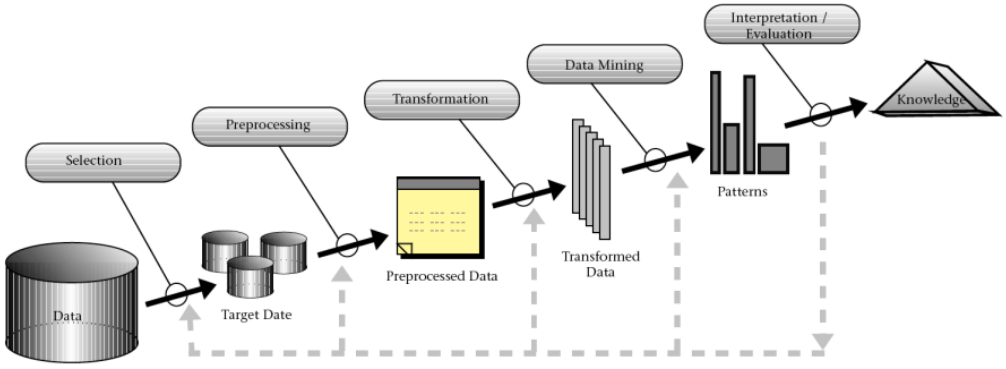
Kelas Pelanggan	Karakteristik
Superstar	- Pelanggan dengan tingkat loyalitas yang tinggi - Memiliki nilai monetary yang tinggi - Memiliki frekuensi yang tinggi - Memiliki nilai transaksi yang tinggi
Golden Customer	- Memiliki nilai monetary tertinggi kedua - Memilki frekuensi yang tinggi - Memiliki nilai transaksi rata - rata yang tinggi
Typical Customer	- Memiliki nilai monetary dan frekuensi rata-rata - Memiliki nilai transaksi rata-rata - Memiliki nilai frekuensi terendah setelah Dorman Customer
Occational customer	- Memiliki recency rendah (memiliki waktu yang lama dengan rentang waktu terakhir kunjungan) - Melakukan pembelanjaan (monetary) dalam jumlah besar.
Everyday Shopper	- Memiliki peningkatan dalam bertransaksi - Melakukan pembelanjaan dalam jumlah kecil - Memiliki nilai transaksi yang rendah - Memiliki frekuensi dan monetary terendah
Dormant Customer	Memiliki waktu yang lama ketika masa terakhir kunjungan (memiliki recency terendah)

2.1.3 Data Mining

Data mining merupakan sebuah aktivitas yang diterapkan pada data berskala besar dengan tujuan untuk mencari pola tersembunyi dan informasi yang dapat digunakan untuk pengambilan keputusan. Menurut Vercellis (2009) data mining adalah aktivitas yang menggambarkan sebuah proses analisis yang terjadi secara iteratif pada database yang besar, dengan tujuan mengekstrak informasi dan knowledge yang akurat dan berpotensial berguna untuk knowledge workers yang berhubungan dengan pengambilan keputusan. Sedangkan menurut Han Kamber & Pei (2006) data mining adalah aktivitas untuk menemukan pola tersembunyi dari sekumpulan data yang berjumlah besar yang tersimpan dalam database, datawarehouse atau media penyimpanan data lainnya.

Data mining merupakan bagian dari *knowldge discovery data* (KDD) yang merupakan proses ekstraksi informasi yang berguna, tidak diketahui sebelumnya, dan tersembunyi dari data (Bramer, 2016). Data mining memiliki beragam metode yang bisa digunakan, diantaranya adalah metode KDD, SEMMA dan CRISP-DM. Setiap proses memiliki metode yang berbeda-beda dalam pencarian informasi penting yang ada di dalam suatu data.

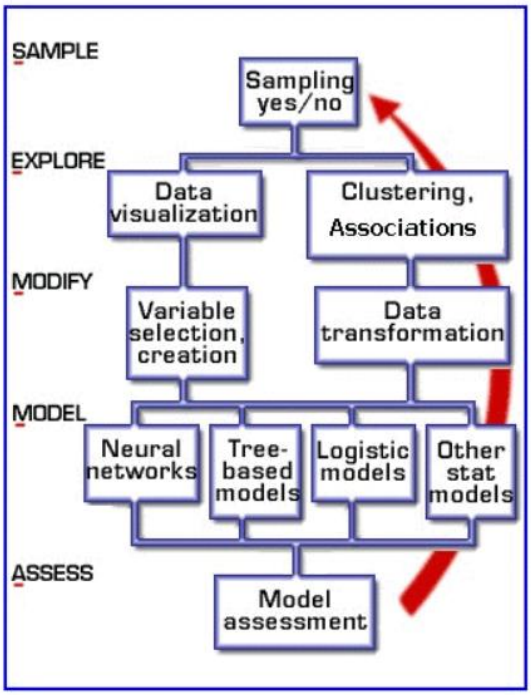
KDD (*Knowledge Discovery in Database*) adalah salah satu metode yang bisa digunakan dalam melakukan data mining. Fayyad, dkk. (1996) mendefinisikan KDD sebagai proses dari menggunakan metode data mining untuk mencari informasi-informasi yang berharga, pola yang ada di dalam data, yang melibatkan algoritma untuk mengidentifikasi pola pada data. Siklus proses KDD dapat dilihat pada Gambar 2.1.



Gambar 2.1 Siklus Proses KDD
(Sumber: Fayyad, U. M., dkk, 1996)

Terdapat lima tahap dalam proses KDD yaitu *selection, preprocessing, transformation, data mining, interpretation/evaluation*. *Selection*, yaitu membuat sebuah target data, fokus dalam bagian dari variabel atau sampel data yang mana *discovery* akan dilakukan. *Preprocessing*, yaitu *cleaning* target data dengan tujuan mendapatkan data yang konsisten. *Transformation*, yaitu transformasi data menggunakan reduksi dimensional atau metode transformasi. Data Mining, yaitu mencari pola menarik di dalam sebuah bentuk tertentu, bergantung dari tujuan data mining (biasanya prediksi). *Interpretation/Evaluation*, yaitu interpretasi dan evaluasi dari pola yang sudah dimining.

SEMMA merupakan singkatan dari *Sample, Explore, Modify, Model,* dan *Assess* yang merupakan tahapan dalam data mining. Metode ini ditemukan oleh SAS *Institute* yang dapat digunakan untuk memudahkan pengguna untuk memprediksi tentang variable-variabel yang mengacu pada proses sebuah proyek data mining. Masing-masing tahapan memiliki peran sendiri dalam proses data mining dan memiliki manfaat dalam proses data mining tersebut. Siklus proses SEMMA dapat dilihat pada Gambar 2.2.

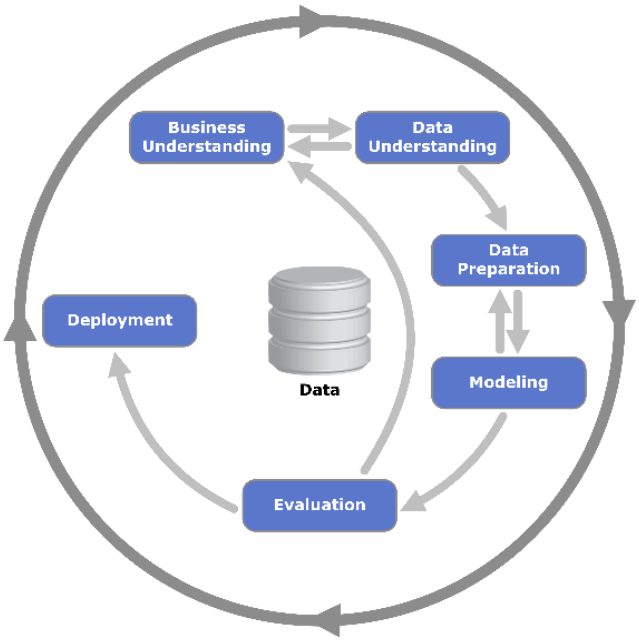


Gambar 2.2 Siklus proses SEMMA
(Sumber: SAS Institute Inc., 2017)

Proses data mining SEMMA memiliki lima proses tahapan yaitu *Sample, Explore, Modify, Model, dan Assess*. *Sample*, yaitu mengambil sampel data. Tahap ini merupakan opsional. *Explore*, yaitu mengeksplorasi data untuk pola dan keanehan yang tidak diharapkan dengan tujuan untuk mendapatkan pengertian dan ide. *Modify*, yaitu memodifikasi data dengan membuat, menyeleksi dan mentransformasi variabel-variabel untuk fokus pada proses pemilihan model. *Model*, yaitu memodelkan data dengan menyediakan software untuk mencari kombinasi data yang memprediksi hasil terpercaya yang diinginkan secara otomatis. *Assess*, yaitu menilai data dengan mengevaluasi kegunaan dan keandalan penemuan dari proses data mining dan mengevaluasi sebaik mana itu bekerja.

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah salah satu model proses data mining (*datamining framework*) yang awalnya pada tahun 1996 dibangun oleh lima perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA. *Framework* ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan

methodology standard non-proprietary bagi data mining. Siklus proses CRISP-DM dapat dilihat pada Gambar 2.3.



Gambar 2.3 Siklus proses CRISP-DM
(Sumber : Chapman dkk, 2000)

Terdapat enam tahap siklus pengembangan data mining (Chapman dkk, 2000) yaitu sebagai berikut :

1) *Business understanding* (pemahaman bisnis)

   Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemahkan pengetahuan ini ke dalam pendefinisian masalah dalam data mining. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2) *Data understanding* (pemahaman data)

   Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3) *Data preparation* (persiapan data)

   Tahap ini meliputi kegiatan untuk membangun kumpulan data akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan.

4) *Modeling* (pemodelan)

   Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal.

5) *Evaluation* (evaluasi)

   Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisis data. Pada tahap ini akan dilakukan evaluasi terhadap apakah model dapat mencapai tujuan yang ditetapkan pada fase awal (pemahaman data).

6) *Deployment* (pengembangan)

   Pada tahap terakhir, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap pengembangan dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses data mining yang berulang dalam perusahaan.

Dari tiga metode data mining tersebut, Mariscal, Marban, and Fernandez (2010) menyatakan bahwa CRISP-DM sebagai *defacto* menjadi standar untuk pengembangan proyek data mining dan *knowledge discovery* karena paling banyak digunakan dalam pengembangan data mining. Hal tersebut diketahui dari hasil survei "Penggunaan Metodologi dalam Proyek Data Mining", yang menunjukkan bahwa CRISP-DM dari tahun ke tahun menjadi metode yang paling banyak digunakan secara luas di kalangan industri, hal tersebut dikarenakan keunggulannya dalam menyelesaikan banyak persoalan dalam proyek proyek data mining. Oleh karena itu, pada penelitian ini metode data mining yang akan digunakan adalah metode CRISP-DM.

Dalam melakukan analisis data mining, terdapat dua pendekatan berdasarkan tugas dan tujuan analisis, yaitu *supervised learning* dan *unsupervised learning* (Vercellis, C, 2009). *Supervised learning* merupakan sebuah proses pengelompokkan data yang telah memiliki label dan akan dimasukkan/dikelompokkan berdasarkan labelnya, juga algoritma yang terdapat pada *supervised* bertujuan untuk memperkirakan atau memprediksi fungsi pada bidang pemetaan sehingga ketika ada variable input (X) maka dapat memprediksi variable output (Y). Sedangkan *unsupervised learning* merupakan sebuah proses pengelompokkan data yang tidak diberi label, tipe algoritma yang memiliki variable input (X) tetapi tidak memiliki variable output yang sesuai. Tujuan dari *unsupervised* adalah untuk memodelkan struktur data agar dapat mempelajari data-data tersebut lebih lanjut lagi, mengindentifikasikan pola – pola dalam sekumpulan data yang pada umumnya tidak diklasifikasikan.

Terdapat beberapa jenis metode analisis data mining dengan *supervised learning* (Vercellis, C, 2009), yaitu diantaranya:

a) *Classification* (klasifikasi), merupakan proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.
b) *Regression* (Regresi), merupakan metode analisis yang memakai hasil observasi masa lalu utk memprediksi nilai atribut target berdasarkan atribut penjelas hasil observasi di masa depan. Klasifikasi dapat dijadikan regresi, dan sebaliknya
c) *Characterization and discrimination* (Karakterisasi dan diskriminasi), merupakan metode analisis untuk memandingkan nilai distribusi dari atribut-atribut yg ada di dalam suatu kelas, dan mendeteksi perbedaan antara suatu kelas dengan kelas lain melalui perbandingan distribusi nilai.
d) *Time series*, merupakan metode analisis untuk menginvestigasi data yang memiliki dinamika waktu dan bertujuan untuk memprediksi nilai atribut target dari satu periode mendatang atau lebih.

Metode analisis data mining dengan *unsupervised learning* (Vercellis, C, 2009), yaitu diantaranya:

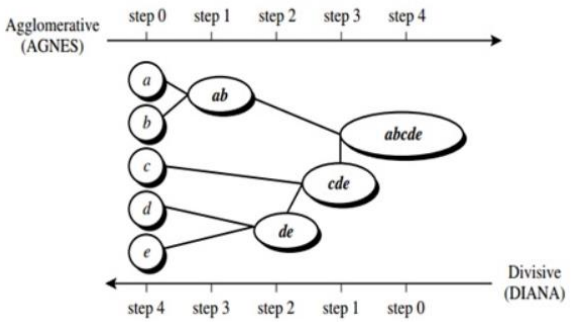
a) *Association* (asosiasi), dinamakan juga analisis keranjang pasar dimana fungsi ini mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain.
b) *Clustering* (pengelompokan), yaitu merupakan motode analisis yang bertujuan untuk melakukan segmentasi populasi yang heterogen menjadi sejumlah kelompok yang beranggotakan observasi dengan karakteristik yang homogen.
c) *Descript ion and visualization* (deskripsi dan visualisasi), merupakan metode analisis untuk memberi gambaran secara ringkas bagi sekumpulan data yang jumlahnya sangat besar sehingga dapat memberikan penjelasan tentang pola yang tersembunyi didalam *dataset* dan mengarah ke pemahaman yang lebih baik tentang fenomena dari *dataset*.

2.1.4 *Clustering*

Clustering merupakan salah satu metode analisis data mining dengan *unsupervised learning*. *Clustering* melakukan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*. Objek yang di dalam *cluster* memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain. Menurut Tan (2006) *clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum.

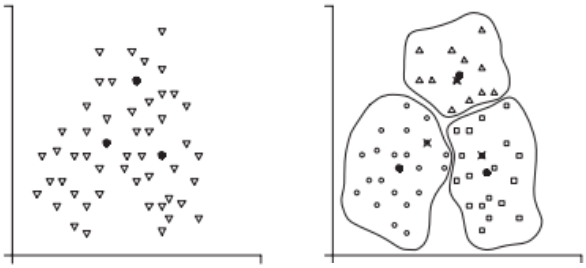
Tujuan dari *clustering* adalah untuk menemukan pengelompokan alami dari serangkaian pola, titik atau objek. Teknik dalam *clustering* harus memenuhi dua kriteria, pertama adalah tiap kelompok adalah homogen; objek yang berada dalam kelompok yang sama pasti mirip satu sama lain, dan yang kedua tiap grup/kelompok harus berbeda dengan kelompok lain atau objek dalam kelompok satu harus berbeda dengan semua kelompok lain (Aggelis, V., & Christodoulakis, D., 2005)

Proses *clustering* tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*. Tan, P.N., Steinbach, M., dan Kumar, V (2006) membagi algoritma *clustering* menjadi dua kelompok yaitu *hierarchical* (hirarki) dan *partitional* (partisi). Pada algoritma *clustering hierarchical* (hirarki), pengelompokan data dilakukan dengan membuat suatu bagan hirarki (dendrogram) dengan tujuan menunjukkan kemiripan antar data. Setiap data yang mirip akan memiliki hubungan hirarki yang dekat dan membentuk *cluster* data. Secara umum, algoritma *clustering* hirarki dibagi menjadi dua jenis yaitu Agglomerative dan Divisive. Agglomerative *clustering* biasa disebut juga sebagai Agglomerative Nesting (AGNES) dimana cara kerja dalam melakukan pengelompokan data menggunakan *bottom-up manner*. Prosesnya dimulai dengan menganggap setiap data sebagai satu *cluster* kecil (*leaf*) yang hanya memiliki satu anggota saja, lalu pada tahap selanjutnya dua *cluster* yang memiliki kemiripan akan dikelompokkan menjadi satu *cluster* yang lebih besar (*nodes*). Proses ini akan dilakukan terus menerus hingga semua data menjadi satu *cluster* besar (*root*). Divisive *clustering* biasa disebut juga sebagai Divisive Analysis (DIANA) dimana cara kerja dalam melakukan pengelompokan data menggunakan *top-down manner*. Prosesnya dimulai dengan menganggap satu set data sebagai satu *cluster* besar (*root*), lalu dalam setiap iterasinya setiap data yang memiliki karakteristik yang berbeda akan dipecah menjadi dua *cluster* yang lebih kecil (*nodes*) dan proses akan terus berjalan hingga setiap data menjadi satu *cluster* kecil (*leaf*) yang hanya memiliki satu anggota saja. Berikut ini gambaran proses hierarchy *clustering* dapat dilihat pada Gambar 2.4.



Gambar 2.4 Gambaran Proses Hierarchy *Clustering*
(Sumber : Han, J., dkk., 2006)

Sedangkan algoritma *clustering partitional* (partisi) menemukan semua *cluster* secara bersamaan sebagai partisi data dan tidak memaksakan struktur hirarki. Pada metode *partitional clustering*, setiap *cluster* memiliki titik pusat *cluster* (*centroid*) dan secara umum metode ini memiliki fungsi tujuan yaitu meminimumkan jarak (*dissimilarity*) dari seluruh data ke pusat *cluster* masing-masing (Jain, A. K, 2009). Contoh algoritma partitional *clustering* diantaranya: K-Means, K-Medoids, Fuzzy K-means dan Mixture Modelling. Berikut ini gambaran proses partitional *clustering* dapat dilihat pada Gambar 2.5.



Gambar 2.5 Gambaran Proses Partitional *Clustering*
(Sumber: Vercellis, C., 2009)

Menurut Han dan Kamber (2006), suatu algoritma *clustering* harus memenuhi syarat-syarat sebagai berikut:

a) Skalabilitas. Suatu metode *clustering* harus mampu menangani data dalam jumlah yang besar. Saat ini data dalam jumlah besar sudah sangat umum digunakan dalam berbagai bidang misalnya saja suatu database. Tidak hanya berisi ratusan objek, suatu database dengan ukuran besar bahkan berisi lebih dari jutaan objek
b) Kemampuan analisa beragam bentuk data. Algortima *clustering* harus mampu dimplementasikan pada berbagai macam bentuk data seperti data nominal, ordinal maupun gabungannya.
c) Menemukan *cluster* dengan bentuk yang tidak terduga. Banyak algoritma *clustering* yang menggunakan metode Euclidean atau Manhattan yang hasilnya berbentuk bulat. Padahal hasil *clustering* dapat berbentuk aneh dan tidak sama antara satu dengan yang lain. Karenanya dibutuhkan kemampuan untuk

menganalisa *cluster* dengan bentuk apapun pada suatu algoritma *clustering*

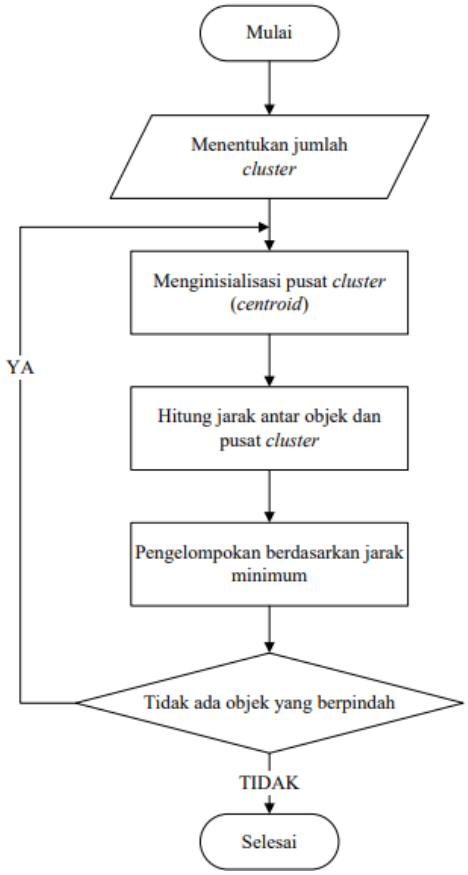
d) Kemampuan untuk dapat menangani *noise*. Data tidak selalu dalam keadaan baik. Ada kalanya terdapat data yang rusak, tidak dimengerti atau hilang. Karena system inilah, suatu algortima *clustering* dituntut untuk mampu menangani data yang rusak.
e) Sensitifitas terhadap perubahan input. Perubahan atau penambahan data pada input dapat menyebabkan terjadi perubahan pada *cluster* yang telah ada bahkan bisa menyebabkan perubahan yang mencolok apabila menggunakan algoritma *clustering* yang memiliki tingkat sensitifitas rendah
f) Mampu melakukan *clustering* untuk data dimensi tinggi. Suatu kelompok data dapat berisi banyak dimensi ataupun atribut. Untuk itu diperlukan algoritma *clustering* yang mampu menangani data dengan dimensi yang jumlahnya tidak sedikit
g) Interpresasi dan kegunaan. Hasil dari *clustering* harus dapat diinterpretasikan dan berguna.

2.1.5 K-Means

Algoritma K-Means merupakan salah satu metode *clustering* partisi yang sangat populer dan banyak dipelajari karena sederhana dan mudah diterapkan (Wei, dkk, 2020; Sheshasaayee; 2018; Jamal, 2019). Algoritma ini pertama kali diusulkan oleh MacQueen (1967) dan dikembangkan oleh Hartigan dan Wong (1975) dengan tujuan untuk dapat membagi M data point dalam N dimensi kedalam sejumlah k *cluster* dimana proses klastering dilakukan dengan meminimalkan jarak *sum squares* antara data dengan masing masing pusat *cluster* (*centroid-based*) (Jain, 2009).

Menurut Larose (2005), algoritma K-Means merupakan sebuah metode sederhana untuk membagi suatu kumpulan data dalam suatu angka spesifik dari *cluster*, yaitu k. K-Means merupakan suatu metode data *clustering* non hirarki yang mempartisi data ke dalam bentuk satu atau lebih *cluster* atau kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam

kelompok lain. Proses *Clustering* meggunakan algoritma K-Means ditunjukkan oleh *flowchart* pada Gambar 2.6.



Gambar 2.6 *Flowchart* Algoritma K-Means
(Sumber: Adiana, B. E., dkk., 2018)

Tahapan Algoritma K-Means:

1) Menentukan jumlah *cluster* yang akan dibentuk
2) Menginisialisasi pusat *cluster* awal (*centroid*) dari masing-masing *cluster* secara acak
3) Menghitung jarak setiap objek atau data observasi terhadap pusat *cluster* (*centroid*) yang dipilih menggunakan rumus *Euclidian distance* pada persamaan (1).

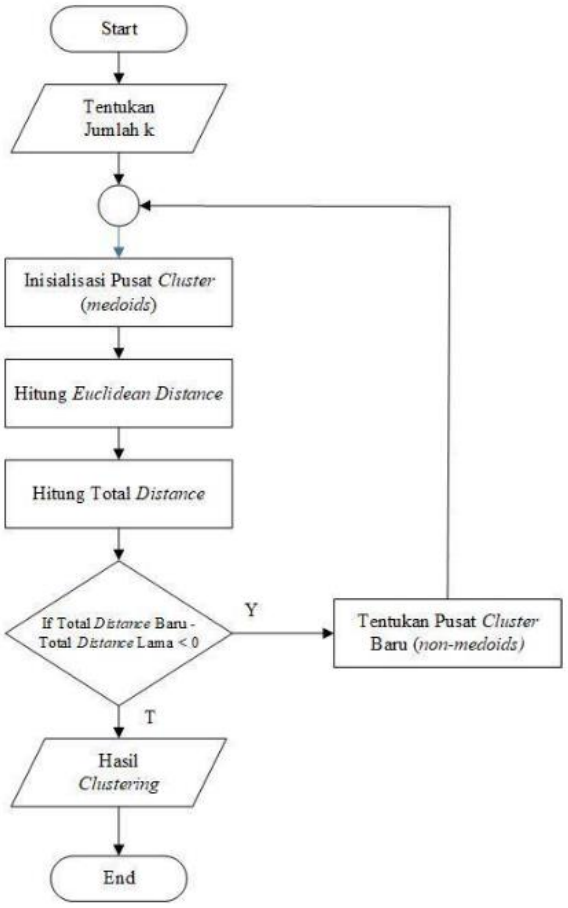
$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

4) Mengelompokkan setiap data observasi berdasarkan jarak minimum atau kedekatannya dengan *centroid* (jarak terkecil)
5) Memperbaharui nilai *centroid* baru yang diperoleh dari rata-rata (*means*) *cluster* yang bersangkutan
6) Melakukan perulangan dari langkah 3 hingga 5, sampai anggota tiap *cluster* tidak ada yang berubah

2.1.6 K-Medoids

Algoritma K-Medoids atau sering disebut juga dengan algoritma PAM (*Partitioning Around Medoid*) merupakan salah satu metode *clustering* partisi yang dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw (1987) . K-Medoids hadir untuk mengatasi kelemahan K-Means yang sensitif terhadap *outlier* karena suatu objek dengan suatu nilai yang besar mungkin secara substansial menyimpang dari distribusi data (Han, J., & Kamber, M, 2006; Atmaja, E. H. S., 2019).

Algoritma K-Medoids telah diusulkan untuk meningkatkan algoritma K-Means dan memecahkan beberapa masalah, termasuk sensitivitas terhadap *outlier*. Dalam beberapa kumpulan data yang memiliki beberapa fitur nominal (non numerik), tidak mungkin untuk memilih pusat *cluster* dengan menggunakan perhitungan rata-rata dari titik-titik di dalam *cluster* (Mousavi, S., Boroujeni, F. Z., & Aryanmehr, S., 2020; Rahman, F, et al , 2020). Oleh karena itu, metode K-Medoids menyelesaikan masalah ini dengan menggunakan *medoid* (perwakilan) data sebagai ganti dari *mean* untuk pusat massa *cluster*. *Medoid* adalah objek data paling sentral di antara titik-titik *cluster*. Oleh karena itu, K objek dipilih secara acak sebagai *medoid* untuk mewakili *cluster*, dan semua objek data ditugaskan ke *cluster* dengan *medoid* terdekat. Setelah memproses semua titik, *medoid* baru ditentukan untuk setiap klaster, yang dapat menjadi perwakilan klaster yang lebih baik, dan dengan demikian, seluruh proses akan diulang. Dalam setiap pengulangan, *medoid* berubah dan algoritma berlanjut sampai *medoid* lainnya tidak berubah (Mousavi, S., Boroujeni, F. Z., & Aryanmehr, S., 2020). Proses *Clustering* meggunakan algoritma K-Medoids ditunjukkan oleh *flowchart* pada Gambar 2.7.



Gambar 2.7 *Flowchart* Algoritma K-Medoids
(Sumber: Andini, A. D., dkk., 2020)

Tahapan Algoritma K-Medoids:

1) Menentukan jumlah *cluster* yang akan dibentuk (nilai k).
2) Menginisialisasi pusat *cluster* awal (*medoids*) dari masing-masing *Cluster* secara acak.
3) Menghitung setiap data observasi (objek) ke *cluster* terdekat menggunakan persamaan ukuran jarak *Euclidian Distance* dengan persamaan (1).
4) Menghitung secara acak objek pada masing-masing *cluster* sebagai kandidat *medoid* baru.
5) Menghitung jarak setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru.
6) Menghitung total simpangan (S) dengan menghitung nilai total *distance* baru – total *distance* lama. Jika S < 0, maka tukar objek dengan data *cluster* untuk

membentuk sekumpulan k objek baru sebagai *medoid*.

7) Mengulangi langkah 4 sampai 6 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing.

2.1.7 Agglomerative

Agglomerative atau sering disebut juga *Agglomerative Nesting* (AGNES) atau *Agglomerative Hierarchical Clustering* (AHC) merupakan salah satu metode *clustering* hirarki yang mengelompokan data secara berulang-ulang berdasarkan tingkat kemiripan data satu sama lain dan membentuk sebuah hirarki (Justitia, R. P., Hidayat, N., & Santoso, E., 2021). Pada teknik pengelompokan ini menggunakan pendekatan *bottom-up* dengan mengelompokan data mulai dari data yang berupa *singleton* atau individu kemudian digabung menjadi satu *cluster*/kelompok secara berulang sehingga seluruh data terkelompokan. Untuk satu set objek data, algoritma pada awalnya memperlakukan setiap objek sebagai satu *cluster*. Kemudian menggabungkan *cluster* menjadi *cluster* yang lebih besar di setiap iterasi. Proses ini berlanjut sampai terbentuk satu *cluster* atau beberapa kondisi yang telah ditentukan terpenuhi (Shihab, S. H., Afroge, S., & Mishu, S. Z., 2019)

Dalam *hierarchical clustering*, proses penggabungan *cluster-cluster* kecil menjadi satu hirarki utuh dilakukan melalui beberapa parameter jarak atau pendekatan *Linkage Method* (Exasanti, D., & Jananto, A., 2021). Berikut ini *linkage method* yang sering digunakan pada Agglomerative:

1) *Single Linkage*.

   Teknik yang menggabungkan *cluster-cluster* menurut jarak antara anggota-anggota terdekat di antara dua *cluster*.

2) *Complete Linkage*

   Teknik yang menggabungkan *cluster-cluster* menurut jarak antara anggota-anggota terjauh di antara dua *cluster*.

3) *Average Linkage*

   Teknik yang menggabungkan *cluster-cluster* menurut jarak rata-rata pasangan anggota masing-masing pada himpunan antara dua *cluster*.

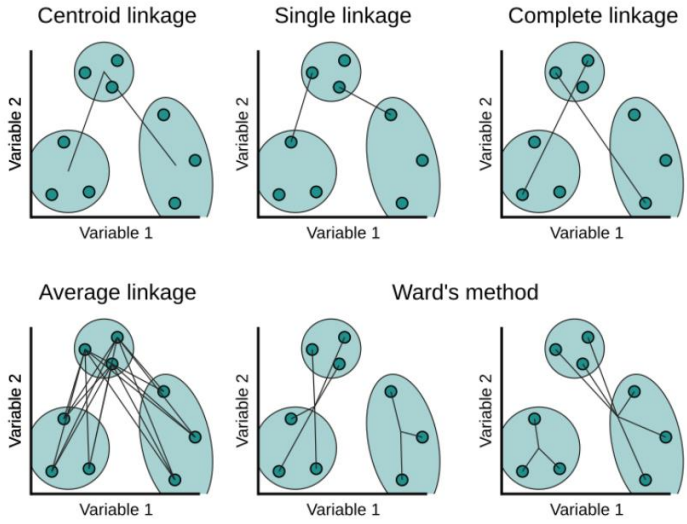
4) *Centroid Linkage*

Teknik yang menggabungkan *cluster-cluster* menurut jarak antar *centroid* pada dua *cluster*. Perhitungan *centroid* disini menggunakan rata-rata pada suatu variabel x. Dendrogram yang terbentuk akan berdasarkan *cluster* dengan jarak antar centroid paling kecil.

5) *Ward's minimum Variance*

Pada metode ini, ditiap iterasinya akan dibentuk *cluster-cluster* yang kemudian dihitung nilai *within sum of square* (WSS) tiap *cluster*. WSS dapat diartikan sebagai jumlah dari jarak tiap observasi ke nilai tengah *cluster*. *Cluster-cluster* yang menghasilkan *within sum of square* terkecil akan diambil kemudian digabungkan hingga membentuk satu dendrogram utuh.

Gambaran *Linkage method* pada *Agglomerative Hierarchical Clustering* dapat dilihat pada Gambar 2.8.



Gambar 2.8 *Linkage method pada Agglomerative Hierarchical Clustering*
(Sumber: Rhys, H., 2020)

Tahapan dalam *Agglomerative Hierarchical Clustering* yaitu:

1) Menghitung matriks jarak menggunakan *Euclidian Distance* dengan persamaan (1).
2) Menggabungkan dua *cluster* terdekat. Jika jarak objek a dengan b memiliki nilai jarak paling kecil dibandingkan jarak antar objek lainnya dalam matriks jarak *Euclidean*, maka gabungan dua *cluster* pada tahap pertama adalah d_ab

3) Perbarui matriks jarak sesuai dengan teknik pengelompokan *Linkage method*. Jika d_ab adalah jarak terdekat dari matriks jarak *Euclidean*, maka rumus untuk metode Agglomerative dapat dilihat pada persamaan (2).

(2)

$$d_{(ab)c} = \min \{d_{a,c}; d_{b,c}\}$$ rumus *single linkage*

$$d_{(ab)c} = \text{average } \{d_{a,c}; d_{b,c}\}$$ rumus *average linkage*

$$d_{(ab)c} = \max \{d_{a,c}; d_{b,c}\}$$ rumus *complete linkage*

4) Ulangi langkah 2 dan 3 sampai hanya tersisa satu *cluster*
5) Membuat dendrogram

2.2 Alur Penelitian (*Roadmap*)

2.2.1 Penelitian Terdahulu

Terdapat beberapa penelitian terdahulu mengenai segmentasi pelanggan dengan model RFM menggunakan metode *clustering* baik partisi maupun hirarki yang menjadi referensi pada penelitian ini, dapat dilihat pada Tabel 2.2.

Tabel 2.2 Perbandingan Dengan Penelitian Terdahulu

No	Nama, Tahun	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian ini
1.	Hung, P. D., Lien, N. T. T., & Ngoc, N. D. (2019)	Customer Segmentation Using Hierarchical Agglomerative *Clustering*	Dari *dataset* kartu kredit diperoleh 3 *cluster* segmentasi pelanggan. Peneliti dapat memprommosikan strategi pemasaran yang tepat yang lebih menguntungkan, namun kelemahan dari metode ini cukup lambat dan tergantung pada perangkat keras.	-Sama mencari segmentasi pelanggan -Sama menggunakan bahasa pemrograman R	- Metode yang digunakan hanya Agglomerative - Atribut yang digunakan: balance, cash advance, purchase trx, dan tenure -Data *outlier* dihilangkan	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative -Atribut menggunakan model RFM -Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan.
2.	Shirole, R., Salokhe, L., & Jadhav, S. (2021).	Customer Segmentation using RFM Model and K-Means *Clustering*	Dari *dataset* e-commerce diperoleh 4 *cluster* pelanggan berdasarkan skor RFM. Hasil tersebut dapat membantu perusahaan untuk mengembangkan strategi	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan hanya K-Means - Data *outlier* dihilangkan	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak

			pasar dan sebagai media promosi kepada pelanggan setia.			dihilangkan - menggunakan bahasa pemrograman R
3.	Shihab, S. H., Afroge, S., & Mishu, S. Z. (2019)	RFM based market segmentation approach using advanced k-means and Agglomerative *clustering*: a comparative study	Metode advanced K-means lebih baik daripada Agglomerative berdasarkan dari waktu prosesnya sebesar 27,8% dan 97,8% serta lebih baik daripada standard K-means berdasarkan jarak intra *cluster* dan jarak inter *cluster*	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang dibandingkan adalah K-Means dan Agglomerative - Data *outlier* tidak dihilangkan	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan - menggunakan bahasa pemrograman R
4.	Aryuni, M., Madyatmadja, E. D., & Miranda, E. (2018)	Customer segmentation in XYZ bank using K-means and K-medoids *clustering*	Metode K-Means mengungguli metode K-Medoids berdasarkan jarak intra *cluster* (AWC). Sementara berdasarkan indeks Davies-Bouldin, kinerja K-Means sedikit	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang dibandingkan adalah K-Means dan K-Medoids - Data *outlier* tidak dihilangkan	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan

			lebih baik daripada K-Medoids			- menggunakan bahasa pemrograman R
5.	Sheshasaayee, A., & Logeshwari, L. (2018)	Implementation of *clustering* technique based RFM analysis for customer behaviour in online transactions	Dari *dataset* belanja online diperoleh 2 *cluster* pelanggan. *Cluster* 0 frekuensi yang lebih banyak dan membutuhkan sedikit konsentrasi dan loyalitas. *Cluster* 1 adalah target pelanggan Karena membutuhkan lebih banyak penawaran dan pengingat iklan terbaik.	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan hanya K-Means - Data *outlier* tidak dibahas	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan - menggunakan bahasa pemrograman R
6.	Monalisa, S., Nadya, P., & Novita, R. (2019).	Analysis for customer lifetime value categorization with RFM model	Pelanggan perusahaan LWC dikategorikan menjadi 3 *cluster* yaitu superstar customer, typical customer dan dormant customer. Perusahaan LWC dapat menjalankan strategi dalam	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan adalah Fuzzy C-Means - Data *outlier* tidak dibahas	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan

			mengelola pelanggannya sesuai dengan jenis portofolionya.			- menggunakan bahasa pemrograman R
7.	Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020)	An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm	Dari *dataset* penjualan online diperoleh 4 *cluster* pelanggan berdasarkan perilaku pembelian. Hasil penelitian mendukung perbaikan dari beberapa indeks kinerja utama seperti pertumbuhan pelanggan aktif, total volume pembelian, dan total jumlah konsumsi.	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan hanya K-Means - Data *outlier* dihilangkan	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan - menggunakan bahasa pemrograman R
8.	Dawane, V., Waghodekar, P., & Pagare, J. (2021)	RFM Analysis Using K-Means *Clustering* to Improve Revenue and	Dari *dataset* distributor ritel diperoleh 4 *cluster* pelanggan. Hasil tersebut dapat membantu merancang strategi pemasaran yang ditargetkan untuk setiap	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan hanya K-Means - Data *outlier* dihilangkan - Menggunakan bahasa	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak

		Customer Retention	segmen pelanggan.		pemrograman phyton	dihilangkan - menggunakan bahasa pemrograman R
9.	Wei, J. T., Lin, S. Y., Yang, Y. Z., & Wu, H. H. (2020)	Using a combination of RFM model and *cluster* analysis to analyze customers' values of a veterinary hospital	Dari *dataset* rumah sakit hewan diperoleh 12 *cluster* pelanggan. Tujuh *cluster* ditemukan sebagai pelanggan terbaik atau loyal. Lima klaster lainnya merupakan pelanggan tidak pasti. Dari hasil tersebut rumah sakit hewan dapat membuat strategi pemasaran yang unik untuk berbagai jenis pelanggan.	-Sama mencari segmentasi pelanggan -Model yang digunakan sama RFM	- Metode yang digunakan adalah kombinasi self-organizing maps (SOM) dan K-means - Data *outlier* tidak dibahas - Menggunakan aplikasi SPSS	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data *outlier* yang dihilangkan dengan yang tidak dihilangkan - menggunakan bahasa pemrograman R
10.	Anitha, P., & Patil, M. M. (2019).	RFM model for customer purchase behavior using	Dengan menggunakan Silhouette Coefficient diperoleh jumlah klaster K = 3 dan K = 5. Segmentasi	-Sama mencari segmentasi pelanggan -Model yang	- Metode yang digunakan hanya K-Means - Data *outlier* tidak	-Membandingkan 3 metode: K-means, K-medoids, Agglomerative - Membandingkan data

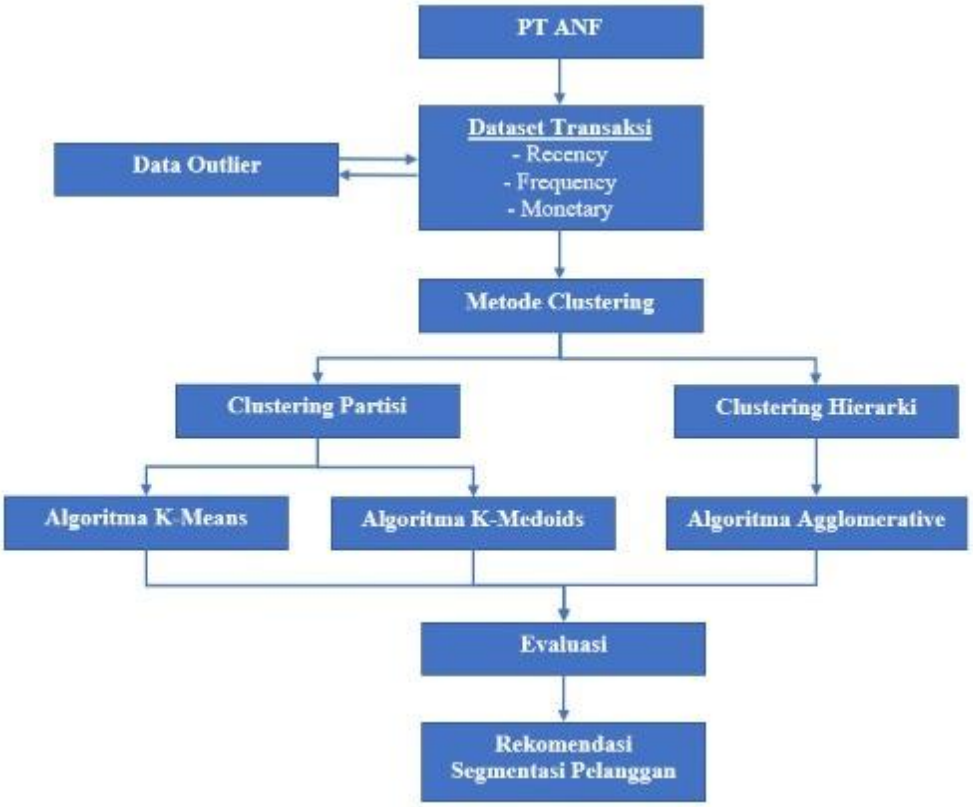
		K-Means algorithm	pelanggan untuk K = 3 lebih optimal daripada K = 5. Dari hasil tersebut dapat memberikan pemahaman yang terorganisir tentang pembelian pelanggan dan pola perilakunya.	digunakan sama RFM	dihilangkan	*outlier* yang dihilangkan dengan yang tidak dihilangkan - menggunakan bahasa pemrograman R

2.2.2 Kerangka Penelitian

Berdasarkan penelitian terdahulu mengenai segmentasi pelanggan dengan model RFM menggunakan metode *clustering* dapat diketahui bahwa sebagian besar peneliti hanya menggunakan satu metode *clustering*, yaitu menggunakan metode *clustering* partisi K-Means dan sisanya menggunakan metode hirarki Agglomerative. Penelitian mengenai segmentasi pelanggan dengan menggunakan metode *clustering* partisi selain K-means (misalnya K-Medoids) dan metode *clustering* hirarki belum banyak dilakukan. Oleh karena itu, penelitian ini berusaha untuk membandingkan segmentasi pelanggan dengan menggunakan metode partisi dalam hal ini K-Means dan K-Medoids dengan metode hirarki Agglomerative.

Pada proses persiapan data, dapat diketahui bahwa terdapat perbedaan dalam menangani data *outlier*, sebagian besar penelitian menghilangkan data *outlier* dan sebagian lain ada yang tidak menghilangkan data *outlier* dan ada juga yang tidak membahas sama sekali mengenai data *outlier*. Penanganan data *outlier* ini perlu diperhatikan karena akan mempengaruhi hasil *clustering* dan rekomendasi strategi mengelola hubungan dengan pelanggan. Oleh karena itu, penelitian ini berusaha untuk membandingkan proses *clustering* dengan menghilangkan data *outlier* dan tidak menghilangkan data *outlier*.

Berdasarkan penjelasan diatas, maka dapat digambarkan kerangka penelitian yang akan dilakukan pada gambar 2.9.



Gambar 2.9 Kerangka Penelitian