BAB II
TINJAUAN PUSTAKA

2.1 Segmentasi Pasar

Segmentasi pasar mempunyai peran penting dalam memenuhi kebutuhan konsumen, yaitu melalui tindakan strategi pemasaran yang tepat, jelas dan terarah dengan produk yang ditawarkan sesuai dengan kebutuhan, perilaku dan karakteristik pelanggan.

Pada tahun 1956, pertama kali yang mengenalkan konsep segmentasi ialah Wendell Smith. Konsep yang dikenalkan yaitu mengenai pengelompokkan konsumen berdasarkan kebutuhannya. Tujuannya yaitu supaya perusahaan dapat menelaah kebutuhan sekelompok konsumen pada produk tertentu, agar upaya dalam pemasaran lebih ekonomis dan efektif [3].

Segmentasi merupakan proses membagi pasar potensial ke dalam kelompok konsumen yang berbeda dan memilih satu segmen atau lebih sebagai pasar target yang akan dicapai dengan *marketing mix* yang berbeda. [26]

Menurut Smith segmentasi pasar merupakan pasar yang *heterogen* (ditandai dengan permintaan yang berbeda) dimana sejumlah pasar *homogen* dalam menanggapi preferensi produk yang berbeda di antara segmen pasar [27].

Dasar segmentasi adalah sekumpulan *variabel* yang digunakan untuk melakukan segmentasi pasar. Pengkategorian basis menjadi *basis* segmentasi umum, yang independen dari produk, layanan atau keadaan dan *basis* spesifik produk, yang terkait dengan produk, layanan, dan/atau keadaan tertentu [27].

Ada 6 kriteria umum untuk menentukan profitabilitas dan efektivitas suatu segmentasi, yaitu [27]:

a. *Identifiability* (Identifikasi), artinya kemampuan untuk mengidentifikasi pelanggan dari setiap segmen berdasarkan *variabel* yang mudah diukur.
b. *Substantiality* (Substansial), yang berarti segmen harus cukup besar sehingga menguntungkan bagi perusahaan untuk dibidik. Bagaimana kelompok kecil dapat dan masih menguntungkan untuk ditargetkan sangat tergantung pada jenis perusahaan yang dipertimbangkan.
c. *Accessibility* (Aksesibilitas) adalah kemampuan perusahaan untuk menjangkau segmen yang dibidik.

d. *Responsiveness* (Daya Tanggap) berarti bahwa setiap segmen harus merespon secara unik upaya pemasaran.
e. *Stability* (Stabilitas) merupakan dimana segmen harus stabil dan tidak berubah dalam perilaku atau komposisi dari waktu ke waktu, setidaknya dalam jangka waktu yang cukup lama untuk dapat mengidentifikasi segmen dan menerapkan strategi pemasaran secara tersegmentasi.
f. *Actionability* (Kemampuan dalam Bertindak) yaitu kemampuan untuk memberikan pedoman bagi pengambilan keputusan.

2.1.1 Tujuan Segmentasi Pasar

Sebuah perusahaan dapat berkembang sesuai perencanaannya dengan melakukan pengelompokkan pasar berdasarkan karakteristiknya. Dengan adanya segmentasi pasar, perusahaan mempunyai tujuan sebagai berikut [28]:

a. Biaya relatif murah atau kecil dan hasil penjualan lebih besar supaya tujuan dari pemasaran lebih efisien dan efektif.
b. Pelayanan yang diberikan kepada pembeli lebih baik.
c. Segmen pasarnya lebih mudah dibedakan.
d. Analisis terhadap pasar mudah dilakukan.

2.1.2 Indikator Segmentasi Pasar

Dalam menyusun segmentasi sebuah pasar, tidak ada satu cara yang khusus. Untuk melihat suatu struktur pada pasar, seorang pemasar harus mencoba *variabel-variabel* segmentasi pasar yang kombinasi, sendiri-sendiri dan berbeda. Berikut segmentasi pasar berdasarkan geografi, demografi, psikografi dan perilaku (*behavioristik*) [28]:

a. Segmentasi pasar berdasarkan geografi.
   Desa, kota, kecamatan, kabupaten, provinsi dan bangsa merupakan segmentasi pasar berdasarkan geografi yang membagi pasar ke dalam unit-unit geografi yang berbeda.
b. Segmentasi pasar berdasarkan demografi.
   Kewarganegaraan, jenjang pendidikan, pekerjaan, pendapatan, jumlah keluarga, jumlah penduduk, jenis kelamin dan usia merupakan segmentasi pasar berdasarkan demografi yang membagi pasar ke dalam kelompok-kelompok

berdasarkan *variabel-variabel* demografi. Dalam membedakan kelompok-kelompok konsumen, *variabel* demografi merupakan dasar paling populer. Alasan *variabel* demografi merupakan dasar paling populer karena selain tingkat pilihan, pemakaian dan keinginan konsumen juga lebih memudahkan untuk mengukur dari kebanyakan *variabel.*

c. Segmentasi pasar berdasarkan psikografi.

   Ciri-ciri kepribadian, gaya hidup dan kelas sosial merupakan segmentasi pasar berdasarkan psikografi dimana para pembeli dibagi ke dalam kelompok yang berbeda-beda. Dalam segmentasi ini, terdapat 2 profil psikografis yang berbeda, yaitu gaya hidup dan kelas sosial. Pada profil gaya hidup, minat masyarakat terhadap bermacam-macam barang dipengaruhi oleh gaya hidupnya. Sedangkan pada profil kelas sosial, alat rumah tangga, pakaian dan mobil menunjukan bahwa suatu masyarakat tertentu di kelas sosial mempunyai pengaruh kuat terhadap pilihan seseorang.

d. Segmentasi pasar berdasarkan perilaku (*behavioristik*).

   Sikap, pengetahuan dan tanggapan atau penggunaan mereka terhadap sebuah produk merupakan segmentasi pasar berdasarkan perilaku (*behavioristik*) dimana para konsumen dibagi ke dalam kelompok-kelompok berdasarkan perilakunya (*behavioristik*). Menurut pemasar, variabel perilaku (*behavioristik*) merupakan titik awal terbaik untuk membuat segmen-segmen pasar.

2.2 Produk

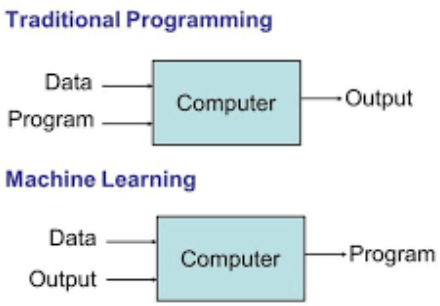
Segala sesuatu yang dapat ditawarkan ke pasar untuk memenuhi kebutuhan atau keinginan konsumen disebut produk. Istilah lain dari produk ialah proses yang memberikan sejumlah nilai kepada konsumen. Sedangkan dalam jasa pendidikan, variasi pilihan, prospek dan reputasi yang ditawarkan kepada pelanggan berupa jasa disebut produk [28].

2.3 *Machine Learning*

Machine learning Pertama kali diperkenalkan oleh Arthur Samuel tahun 1959, dalam jurnalnya yang berjudul "*Some Studies in Machine Learning Using the Game of Checkers*" (IBM *Journal of Research and Development*). Untuk menggantikan manusia

dalam mengambil keputusan, maka digunakanlah *machine learning*. Pengertian *machine learning* sendiri adalah pemrograman komputer untuk mencapai kriteria/performa tertentu dengan menggunakan sekumpulan data training atau pengalaman di masa lalu (*past experience*) [8].

Sebuah sistem yang mampu belajar sendiri untuk memutuskan sesuatu tanpa harus berulang kali diprogram oleh manusia merupakan fokus dari *machine learning.* Dalam pengambilan keputusan dan bisa beradaptasi dengan perubahan yang terjadi merupakan metode dari *machine* atau mesin. Untuk menemukan pola dalam *machine learning,* perlu melakukan analisis pada kumpulan data yang besar [29]. Ada perbedaan antara pemrograman tradisional dengan *machine learning* seperti pada Gambar 2.1.



Gambar 2. 1 Perbedaan Pemrograman Tradisional dan *Machine Learning* [30].

Pada gambar 2.1 menjelaskan bahwa pada pemrograman tradisional program dan data dijalankan pada komputer untuk menghasilkan *output*, sedangkan pemrograman menggunakan *machine learning output* dan data dijalankan pada komputer untuk membuat program kemudian program tersebut bisa digunakan dalam pemrograman tradisional [29].

Data diperlukan oleh *machine learning* dalam mempelajari teori supaya komputer mampu “belajar”. Berbagai ilmu disiplin telah dilibatkan oleh *machine learning*, seperti neurologi, matematika, statistika, dan ilmu komputer [8]. Dalam mengolah data, pengerjaan *machine learning* tidak digunakan sebagai informasi tetapi dikhususkan untuk pengetahuan karena tugas *machine learning* yaitu selain menampilkan data sebagai informasi juga dapat menambang data untuk diekstraksikan pengetahuan dari data tersebut [7].

2.3.1 Pembelajaran Konsep Pada Algoritma *Machine Learning*

Dalam konteks *machine learning*, data training yang digunakan sebagai input oleh algoritma *machine learning* dengan melakukan *concept learning* (pembelajaran konsep) merupakan *training model* atau proses pembelajaran. Sebuah proses yang bertujuan untuk membuat beberapa konsep atau *generalisasi* satu disebut pembelajaran konsep [31].

Didalam pembelajaran konsep, terdapat 2 konsep yaitu label data dan data training. Label data berfungsi sebagai keanggotaan dari konsep tersebut sedangkan data *training* berfungsi sebagai contoh konsep. Contoh dari pembelajaran konsep yang dilakukan oleh algoritma *machine learning* yaitu pola transaksi keuangan kartu debit yang dikelompokkan sebagai transaksi normal atau *fraud* [31]. Pembelajaran konsep dengan algoritma *machine learning* dapat dibagi ke dalam 3 konsep, yaitu [31]:

1. Representasi.

Setiap transaksi keuangan direpresentasikan dengan 6 buah fitur yaitu *amount* (nilai) transaksi pada berbagai waktu sebagai berikut:

$$< x_1, x_2, x_3, x_4, x_5, x_6 >$$

Dimana:

x_1 : nilai transaksi saat ini.

x_2 : rata-rata nilai transaksi kemarin.

x_3 : rata-rata nilai transaksi dua hari lalu.

x_4 : rata-rata nilai transaksi seminggu lalu.

x_5 : rata-rata nilai transaksi dua minggu lalu.

x_6 : rata-rata nilai transaksi sebulan yang lalu.

Setiap sampel transaksi diberi label (y), apabila suatu transaksi dikelompokkan sebagai transaksi *fraud* maka diberi label y = 1 sedangkan apabila suatu transaksi dikelompokkan sebagai transaksi normal maka diberi label y = 0.

Data training terdiri dari n buah contoh transaksi kartu debit yang direpresentasikan dengan sebuah matriks X berukuran $n \times 6$, dimana setiap kolom merepresentasikan *variabel* transaksi dan setiap baris merepresentasikan sebuah transaksi.

2. Optimisasi.

Pemilihan sebuah fungsi $h \in H$ (selanjutnya disebut hipotesis) yang paling optimum dari seluruh kemungkinan hipotesis untuk mengaproksimasi y sehingga $h(x) = \hat{f}(x) = y$ untuk $x \in X$. Untuk mendapatkan nilai maksimal, pada pemilihan parameter model yang optimal dilakukan dengan penelusuran secara sistematik terhadap bidang kurva fungsi objektif didalam ruang multidimensi.

3. Evaluasi.

Algoritma pembelajaran mengevaluasi sejumlah alternatif model yang masing-masing dibangkitkan oleh sejumlah alternatif parameter model merupakan proses optimisasi fungsi objektif. Alat pengukur seberapa optimum sebuah model dibandingkan dengan alternatif model yang lain merupakan fungsi dari objektif.

2.3.2 Tahap-Tahap Membangun *Machine Learning*

Machine learning dibangun melalui beberapa tahap sebagai berikut [29]:

1. Pengumpulan Data.

   Media cetak dan internet merupakan contoh sumber informasi yang dapat dilakukan pada pengumpulan data. Data yang disebar luaskan secara bebas ke publik merupakan data yang dapat dikumpulkan.

2. Mempersiapkan Data Masukan.

   Data masukan yang sesuai dengan format yang dibutuhkan untuk analisis merupakan data masukan yang harus disiapkan.

3. Menganalisis Data Masukan.

   Memisahkan data berdasarkan dimensi masing-masing data dan melihat pola data merupakan proses dalam menganalisis data masukan.

4. Mengikutsertakan Keterlibatan Manusia.

   Untuk meyakinkan bahwa tidak ada *garbage* atau data yang tidak berguna dalam proses *machine learing*, data *training* atau data latih yang digunakan memerlukan keterlibatan manusia.

5. Melatih Algoritma.

   Pada algoritma atau model, pengguna memberikan data yang berkualitas sebagai data latih untuk diproses oleh algoritma dalam mengolah data menjadi informasi serta menyimpannya.

6. Menguji Algoritma.

Pada proses pengujian algoritma yang dilakukan dengan menggunakan data uji perlu dilihat seberapa baik kualitas algoritma yang telah dilatih serta efisiensi dari algoritma tersebut.

7. Menggunakan model.

   Untuk dapat melakukan suatu hal, tahap ini merupakan tahap terakhir untuk algoritma yang diterapkan dalam suatu program.

2.3.3 Perancangan Sistem Berbasis *Machine Learning*

Komponen utama dalam merancang sebuah sistem yang menerapkan *machine learning* sebagai berikut [31]:

1. Pada pengalaman training, ada 2 jenis sifat alternatif *training* model *machine learning* yaitu:
   a. Bersifat langsung, dilakukan oleh algoritma pembelajaran.
   b. *Transfer learning* (bersifat tidak langsung), dimana mempelajari hasil tahapan pembelajaran yang telah dilakukan oleh model lain dengan hasil pembelajaran yang baik.
2. Pada fungsi objektif sebagai dasar proses optimisasi, c*ross entropy* atau *mean square error* merupakan beberapa fungsi objektif yang bisa digunakan didalam *training* model *machine learning.*
3. Pada representasi dari model pembelajaran, model *artificial neural network,* linier dan *probabilistik* merupakan beberapa alternatif representasi model pembelajaran.
4. Pada algoritma pembelajaran, *stohastic gradient descent* merupakan beberapa algoritma pembelajaran yang dapat digunakan untuk menelusuri parameter model dengan mengoptimalkan fungsi objek. Memprediksi *variabel* target dari data *training* dan meng*update* parameter model merupakan langkah penting dari algoritma pembelajaran pada langkah iteratif.

2.3.4 Implementasi *Machine Learning*

Penerapan *machine learning* dalam bidang teknologi saat ini sangat banyak. Berikut beberapa implementasi dari *machine learning* yang digunakan dalam bidang teknologi:

a. *Search Engine*

Search engine adalah program yang dapat mencari informasi pada *database* berdasarkan *keywords*/karakter yang ditentukan oleh *user*. Istilah lain dari *search engine* yaitu mesin pencarian. Yahoo, Google dan Alta Vista merupakan bagian dari *search engine* dimana cara kerjanya dengan memanfaatkan teknologi *machine learning* [32]. Berikut beberapa penerapan *machine learning* pada *search engine* [32]:

1. Pemahaman *query* atau *query understanding*, arti dari *query* ialah *keyword* yang di input oleh pengguna pada halaman *search engine. User intent* adalah tujuan yang dimaksud oleh *user* (melalui *query* tersebut).
2. Klasifikasi *query*, *search engine* menjalankan beberapa algoritma *classifier* untuk mendeteksi berbagai *query*, seperti *shopping query, transactional query, informational query, navigational query.*
3. Pengelompokkan *user*, *search engine* dapat menemukan apa yang dicari dan menampilkan hasilnya berdasarkan kategori *user* yang berbeda-beda.
4. Entitas, *search engine* mendeteksi waktu, lokasi, orang yang ada pada dokumen atau halaman dan mencoba untuk menggambarkan hubungan antar entitas tersebut.
5. Klasifikasi halaman, *search engine* harus dapat memahami jenis dokumen dan jenis halaman yang sedang dicari. Contoh pada jenis dokumen yaitu: tutorial, jurnal, *paper*, *user guide,* sedangkan contoh pada jenis halaman yaitu: *merchant, news,* forum, blog.

b. Sistem rekomendasi, saat seseorang sedang mencari produk berupa pakaian di sebuah situs maka situs tersebut akan muncul daftar rekomendasi pakaian yang dianggap cocok dengan kebutuhannya. Situs semacam ini dapat memunculkan rekomendasi lintas *platform,* mulai dari tablet, *handphone*, laptop, komputer. berikut beberapa layanan *cloud* yang sudah menerapkan sistem rekomendasi berdasarkan aktifitas pencarian, seperti Google, Amazon, Netflix. Contoh lain yang sudah menerapkan sistem rekomendasi, seperti *robot locomotion, stock market analysis, medical diagnosis* [32].

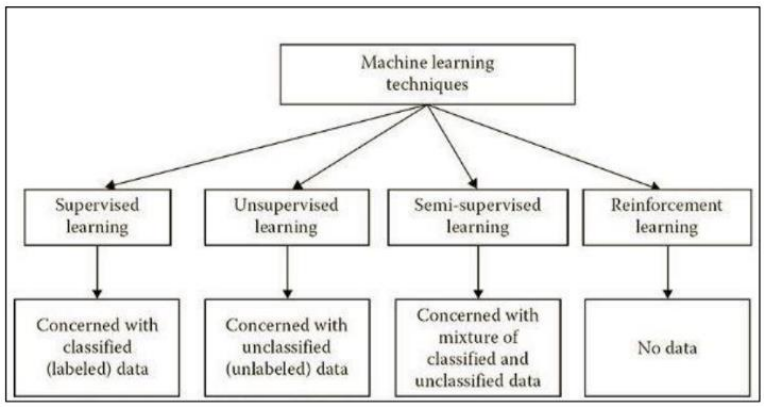
c. Marketing, seperti memprediksi minat seorang calon pembeli terhadap sebuah produk berdasarkan catatan pembelian yang telah dilakukan oleh pelanggan tersebut [31].
d. Perbankan, seperti mengenali apakah *fraud*/transaksi normal merupakan kategori pada sebuah transaksi keuangan [31].
e. Cuaca, seperti memprediksi kelembaban sebuah tempat, kecepatan angin, dan intensitas curah hujan berdasarkan data cuaca tempat tersebut sebelumnya [31].

2.3.5 Tipe-Tipe Algoritma *Machine Learning*

Teknik pembelajaran pada *machine learning* dikategorikan menjadi empat tipe . Keempat tipe tersebut yaitu [29]:

1. *Supervised learning.*
2. *Unsupervised learning.*
3. *Semi supervised learning.*
4. *Reinforcement learning.*

Berikut gambaran mengenai teknik pembelajaran pada *machine learning* yang dikategorikan menjadi empat tipe.



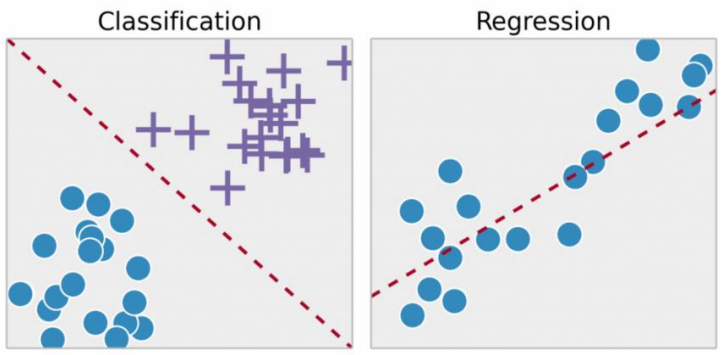
Gambar 2.2 Teknik Pembelajaran Pada *Machine Learning* [33].

Berikut penjelasan mengenai teknik pembelajaran pada *machine learning* yang dikategorikan menjadi empat tipe.

a. *Supervised Learning.*

Machine learning dengan teknik *supervised learning* menggunakan data latih dalam melakukan pembelajaran. Data yang sudah berlabel merupakan data latih pada *supervised learning.* Mampu mengidentifikasi label *input* yang baru dengan

menggunakan *fitur* yang ada untuk melakukan klasifikasi maupun prediksi merupakan tujuan dari *supervised learning* [29].



Gambar 2.3 Ilustrasi Permasalahan *Supervised Learning* [34].

Untuk menyelesaikan komputasi, *supervised learning* biasanya dikelompokkan ke dalam 2 kategori, yaitu:

1. *Classification* (klasifikasi).

   Mengelompokkan dataset berdasarkan kategori tertentu merupakan permasalahan dari *classification* (klasifikasi). Contoh: tinggi/rendah, sakit/sehat [35].

2. *Regression* (regresi).

   Memprediksi sesuatu di masa depan atau *future* merupakan permasalahan dari *regression* (regresi). Contoh: berapa persentase kenaikan harga rumah tiga tahun mendatang [35].

   Dalam *supervised learning*, beberapa algoritma yang sudah dikembangkan di antaranya [8]:

   1. *Decision Trees.*
   2. *The Bayes Classifier.*
   3. *Linier regression.*

b. *Unsupervised Learning.*

   Unsupervised learning merupakan 19 ubsta *machine learning* tidak menggunakan data latih dalam melakukan pembelajaran [29]. Data yang tidak berlabel merupakan data latih pada *unsupervised learning.* Mengelompokkan objek yang 19 ubsta sama dalam suatu area tertentu dan tidak memiliki *19ubstant* target merupakan tujuan dari *unsupervised learning* [29].

Untuk menyelesaikan komputasi, *unsupervised learning* biasanya dikelompokkan ke dalam 3 kategori, yaitu [35]:

1. Asosiasi (*Association*).

   Menemukan relasi antara *20ubstant* dan atribut yang berbeda merupakan tujuan dari *association.* Relasi ditentukan berdasarkan *probabilitas* (peluang) *co-occurrence* (keterkaitan) dari item-item dalam sebuah 20ubstant. *Market-basket analysis* sering digunakan oleh asosiasi (*association*). Contoh: jika *customer* membeli the celup, maka *customer* akan membeli gula pasir [35].

2. Klasterisasi (*clustering*).

   Mengelompokkan sampel dalam *cluster* yang sama berdasarkan *similarity* (kemiripan) merupakan tujuan dari klasterisasi (*clustering*). Contoh: anak-anak yang masih sekolah SMA dapat dikelompokkan sebagai remaja. (Benarkah semua anak SMA adalah remaja?) [35].

3. *Dimensionality Reduction.*

   Mengurangi sejumlah *20ubstant* dari *dataset* namun tetap memastikan informasi yang penting tetap tersedia merupakan arti dari *dimensionality reduction* [35]. Berikut metode yang dapat digunakan untuk mewujudkan *dimensionality reduction* [35]:

   a. *Feature Selection.*

      Memilih *subset* (20ubstant saja) dari *original variables* (20ubstant asal).

   b. *Feature Extraction.*

      Melakukan transformasi data dari *a high dimensional space* (dimensi tinggi) ke *a low dimensional space* (dimensi yang lebih rendah). Dalam *unsupervised learning,* beberapa algoritma yang dapat digunakan adalah sebagai berikut:

      1. *K-Means.*
      2. *K-Medoids.*
      3. *Hierarchical Clustering.*

   c. *Semi Supervised.*

      Semi supervised merupakan kombinasi antara 20ubsta *supervised learning* dan *unsupervised learning.* Data berlabel dan data tidak berlabel yang digunakan secara bersamaan merupakan data latih pada *semi supervised* [29].

   d. *Reinforcement Learning.*

Teknik *reinforcement learning* mengajarkan bagaimana cara bertindak untuk menghadapi suatu masalah, 21ubsta 21ubstant tersebut mempunyai dampak. Teknik *reinforcement learning* diterapkan pada agen cerdas agar dapat menyesuaikan kondisi dilingkungannya [29].

2.4 *Clustering*

Clustering merupakan salah satu analisis *machine learning* dengan metode pembelajaran *unsupervised learning*, 21ubsta dalam pembelajarannya tidak ada atribut 21ubstan yang digunakan sehingga semua atribut masukan dianggap sama [9]. *Clustering* adalah proses menemukan *cluster* yang berisi pengetahuan dan dapat menggambarkan suatu keadaan tertentu dalam suatu *dataset* dengan cara mengelompokkan satu set objek data ke dalam beberapa *cluster* berdasarkan kemiripan objek data sehingga objek dalam sebuah *cluster* memiliki kesamaan tinggi, tetapi sangat berbeda dengan objek di kelompok *cluster* lain [11].

Cluster merupakan proses partisi objek data yang sama ke dalam himpunan bagian. Setiap *cluster* termasuk objek data 21ubsta memiliki kemiripan karakteristik satu sama lain dan berbeda dengan *cluster* yang lain [12]. Kemiripan dan perbedaan objek data dinilai berdasarkan nilai atribut yang menggambarkan objek data biasanya melibatkan pengukuran jarak [13], sehingga pada suatu *dataset* metode pengelompokkan yang berbeda dapat menghasilkan pengelompokkan yang berbeda. Menemukan pengelompokkan alami dari serangkaian objek, titik atau pola merupakan tujuan dari *clustering* [4].

2.4.1 Kegunaan *Clustering*

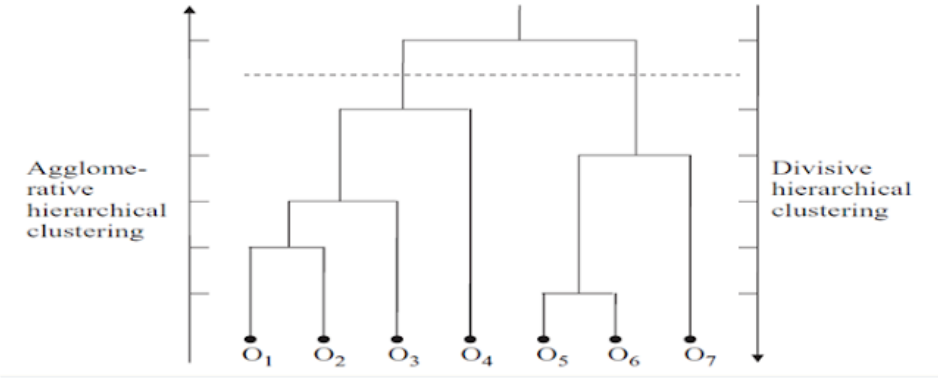
Clustering berguna untuk menemukan kelompok atau *group* yang tidak dikenal dalam data [12]. *Clustering* sebagai pengelompokkan data dapat digunakan untuk menganalisa data 21ubstanti seperti pengenalan pola, *machine learning*, data mining, *bioinformatika*, dan analisis citra [14].

2.4.2 Metode *Clustering*

Secara umum, metode *clustering* dibagi menjadi dua yaitu:

a. Metode Hirarki Klastering (*Hierarchical Clustering Methods*)

Metode hirarki klastering dibagi menjadi dua, yaitu metode penggabungan (*agglomerative*) dan metode pembagian (*divisive*). Pada metode penggabungan (*agglomerative*) dimulai dengan cara menggabungkan setiap objek hingga menjadi satu kelompok, sedangkan pada metode pembagian (*divisive*) dimulai dengan cara memisahkan semua objek pada sebuah kelompok besar menjadi kelompok yang hanya memiliki satu objek. Hasil pengelompokkan dari metode ini disajikan dalam bentuk *dendogram*, yang mirip dengan struktur *tree diagram* atau *diagram* pohon [37]. Contoh metode hirarki klastering diantaranya: *single linkage, complete linkage, average linkage* dan metode *ward* [38]. Berikut 22ubstant mengenai hasil pengelompokkan metode hirarki klastering berupa *dendogram* dapat dilihat pada Gambar 2.4.



Gambar 2.4 Dendogram hasil pengelompokan hirarki klastering (*hierarchical clustering*) [39].

b. Metode Partisi Klastering (*Partitioning Clustering Methods*)

Proses pengelompokkan pada metode partisi dimulai dengan cara pemilihan titik pusat terlebih dahulu dan menentukan jumlah *cluster* sedangkan proses pengelompokkan pada metode hirarki dimulai secara terstruktur atau bertahap tetapi banyaknya *cluster* belum diketahui. Metode partisi klastering mempunyai tujuan yaitu meminimumkan *dissimilarity* (jarak) dari seluruh data ke pusat *cluster* masing-masing [40]. Contoh metode partisi klastering diantaranya: *k-means clustering* dan *k-medoids clustering.*

2.5 *K-Means*

K-means clustering bertujuan untuk membagi n observasi menjadi *k* cluster yang observasinya termasuk dalam *cluster* dengan *mean* terdekat. Algoritma ini bersifat iteratif dan inisialisasi acak dari *centroid* pada awal algoritma [41].

Menganalisis data dengan sistem partisi dan *unsupervised* merupakan metode dari *k-means.* Metode saat ini yang populer untuk *clustering* ialah *k-means*. Algoritma pengklasteran yang paling sederhana yaitu *k-means* dibanding algoritma pengklasteran yang lain. Pengklasteran secara partisi yang memisahkan data ke dalam kelompok yang berbeda merupakan metode *k-means* [16]. Meminimalkan kemiripan data antar *cluster* dan memaksimalkan kemiripan data dalam satu *cluster* pada pengelompokkan data merupakan tujuan dari *k-means*. Fungsi jarak adalah ukuran kemiripan data antar *cluster*. Pemaksimalan kemiripan data didapatkan berdasarkan jarak terpendek antara data terhadap titik *centroid* (titik pusat) [42].

Algoritma *k-means* memiliki nilai *k* dan inputan data dengan arti objek atau data akan didistribusikan ke dalam *k cluster* dan setiap *cluster* memiliki nilai *centroid* (titik pusat) yang menggambarkan *cluster* tersebut [43]. Berdasarkan kedekatan pengamatan nilai rata-rata *cluster*, tiap data kemudian ditetapkan pada setiap pengamatan *cluster*. Pada proses awal, nilai rata-rata pada *cluster* kemudian dihitung secara berulang [15]. Dibutuhkan jumlah *cluster* awal yang diinginkan sebagai masukan dan menghasilkan jumlah *cluster* akhir sebagai *output*. *K* awal dan *k* akhir diperlukan oleh algoritma untuk menghasilkan *cluster k*. Pola *k* sebagai titik awal *centroid* secara acak akan dipilih oleh metode *k-means*. Jika posisi *centroid* baru tidak berubah, maka jumlah iterasi untuk mencapai *cluster centroid* akan dipengaruhi oleh calon *cluster centroid* awal secara *random*. Dengan menggunakan rumus *euclidean distance* yaitu mencari jarak terdekat antara titik *centroid* dengan objek/data, maka nilai *k* yang dipilih menjadi pusat awal bisa dihitung. Sebuah *cluster* dibentuk dengan menggunakan data yang memiliki jarak pendek atau terdekat dengan *centroid* [14]. Pada setiap *cluster* terdapat *centroid* (titik pusat) yang mempresentasikan *cluster* tersebut [45]. Inisialisasi *centroid* dan parameter *k* sangat mempengaruhi hasil *k-means* [9].

2.5.1 Kelebihan *K-Means*

1. Umum digunakan, mudah diimplementasikan, komputasi dan waktu perhitungan yang relatif lebih cepat.
2. Terkenal sederhana dan mudah dipelajari sebagai pemecah masalah pengelompokan dari sebuah *dataset* [15].

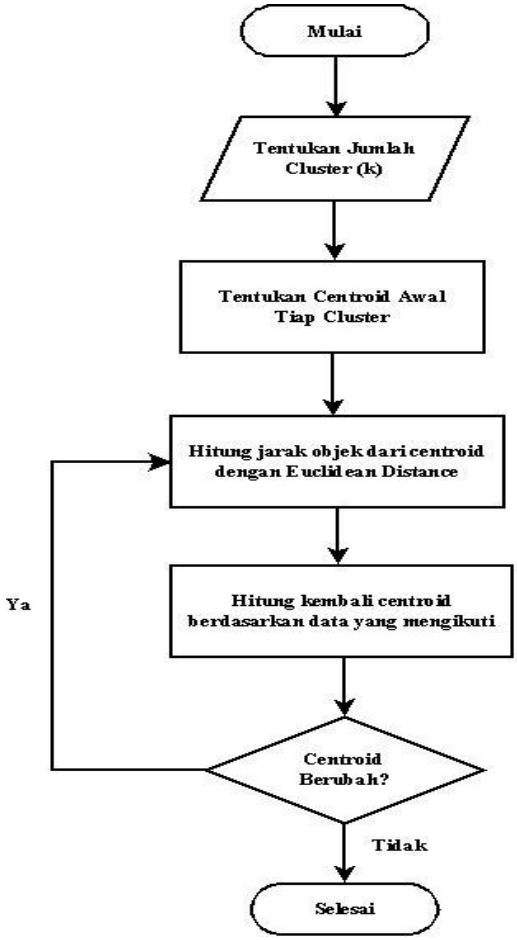
3. Mampu menangani data *outlier* dan memiliki kemampuan untuk mengklaster data yang besar [45].

2.5.2 Kekurangan *K-Means*

1. Hasil *clustering* yang berbeda-beda merupakan proses *clustering* yang dilakukan secara random, susah di*visualisasi* jika atribut yang digunakan terlalu banyak dan sulitnya menentukan nilai *k* yang tepat agar hasil *clustering* optimal [43].
2. Mudah dalam menentukan *centroid* awal apabila jumlah data tidak terlalu banyak, semua atribut dianggap mempunyai bobot yang sama dikarenakan kontribusi dari atribut dalam pengelompokan tidak diketahui, *real cluster* tidak pernah diketahui apabila menggunakan data yang sama dan sebelum dilakukan perhitungan, jumlah *cluster* sebanyak *k* harus ditentukan [44].
3. Sensitif terhadap *noise* dan titik data *outlier* karena nilai rata-rata dapat secara *substantial* dipengaruhi oleh data tersebut [45].

2.5.3 *Flowchart K-Means*

Berikut *flowchart* mengenai proses klastering menggunakan algoritma *k-means* dapat dilihat pada Gambar 2.5.



Gambar 2.5 *Flowchart* Algoritma *K-Means* [46].

2.5.4 Tahapan Perhitungan *K-Means*

Berikut tahapan perhitungan dengan menggunakan *k-means* [46].

1. Menentukan *k* sebagai jumlah *cluster* yang akan dibentuk dimana dalam menentukan banyaknya *cluster k* sesuai dengan tujuan dari pengelompokkan baik secara teoritis maupun konseptual.
2. Menentukan *centroid* (titik pusat) awal dari setiap *cluster* secara acak/*random.*
3. Mengalokasikan semua data ke *centroid* (titik pusat) terdekat dengan matrik jarak yang sudah ditetapkan dengan cara menghitung jarak antara objek dengan *centroid* (titik pusat) menggunakan rumus *euclidean distance* dengan persamaan (1) seperti berikut :

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad ; i = 1,2,3 \ldots n \tag{1}$$

x_i : objek *x* ke-*i*.

y_i : daya *y* ke-*i*.

n : banyaknya objek.

4. Menghitung kembali *centroid* (titik pusat) berdasarkan data *cluster* masing-masing dengan cara menghitung *centroid* (titik pusat) *cluster* ke-*i* berikutnya menggunakan rumus persamaan (2) seperti berikut ini:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \qquad ; i = 1,2,3 \ldots n \tag{2}$$

dimana:

v : *centroid* pada *cluster*.

x_i : objek ke-*i*.

n : banyaknya objek.

5. Melakukan perulangan dari langkah 3 dan 4, sampai posisi *centroid* (titik pusat) tidak berubah.

2.6 *K-Medoids*

Dasar dari algoritma *k-medoids* adalah menemukan *k* objek representatif yang mewakili berbagai atribut struktural data. Oleh karena itu, metode ini disebut sebagai *k-medoids* atau PAM (*Partitioning Around Medoids*). Objek representatif dalam algoritma disebut sebagai *medoid* dan objek tersebut adalah titik terdekat ke pusat. Algoritma *k-medoids* paling umum digunakan dan dikembangkan oleh Kaufman dan Rousseeuw pada tahun 1987 [47].

K-medoids atau PAM (*Partitioning Around Medoids*) adalah algoritma *clustering* yang mirip dengan *k-means.* Kedua algoritma ini mempunyai perbedaan, dimana algoritma *k-medoids* atau PAM (*Partitioning Around Medoids*) menggunakan objek sebagai *medoid* (perwakilan) sebagai pusat *cluster* untuk setiap *cluster*, sedangkan algoritma *k-means* menggunakan nilai *mean* (rata-rata) sebagai pusat *cluster* [17]. Algoritma *k-medoids* lebih baik dibandingkan dengan algoritma *k-means*, dimana *k-medoids* menemukan *k* sebagai objek yang representatif untuk meminimalkan jumlah

ketidaksamaan objek data sedangkan *k-means* menggunakan jumlah jarak *euclidean distance* untuk objek data [18].

Langkah awal *k-medoids* adalah mencari titik yang paling representatif (*medoids*) dalam sebuah *dataset* dengan menghitung jarak dalam kelompok dari semua kemungkinan kombinasi dari *medoids*, sehingga jarak antar titik dalam suatu *cluster* kecil sedangkan jarak titik antar *cluster* besar [19]. K-*medoids* memiliki karakteristik dimana pusat *cluster* berada di antara titik-titik datanya [48].

Untuk mengklasterkan sekumpulan *n* objek menjadi sejumlah *k cluster*, *k-medoids clustering* menggunakan metode pengklasteran partisi. Algoritma *k-medoids* menggunakan objek pada kumpulan objek yang mewakili sebuah *cluster*. *Medoids* adalah objek yang mewakili sebuah *cluster*. *Cluster* dibangun dengan cara menghitung kedekatan yang dimiliki antara *medoids* dengan objek *non medoids* [61]. Setiap objek yang lebih dekat dengan pusat *cluster* akan dikelompokkan dan membentuk *cluster* baru. Algoritma *k-medoids* secara acak menentukan *cluster center* baru dari setiap *cluster* yang terbentuk sebelumnya dan menghitung ulang jarak antara objek dan pusat *cluster* baru yang dihasilkan [4].

2.6.1 Kelebihan *K-Medoids*

1. Mengatasi kelemahan pada algoritma *k-means* yang *sensitive* terhadap *outlier* dan *noise*, dimana objek dengan nilai yang besar memungkinkan menyimpang dari distribusi data serta hasil proses *clustering* tidak bergantung pada urutan masuk *dataset* [17]
2. Pusat *cluster* pada *k-medoids* adalah suatu data yang secara riil berada di tengah *cluster* (*median*), posisi *medoid* juga tidak terlalu terpengaruh terhadap adanya *noise* ataupun distribusi jarak yang berbeda-beda antara data dengan pusat *cluster* dan tidak *sensitive* terhadap *outlier* [43].
3. Cukup efisien untuk *dataset* yang kecil [19].

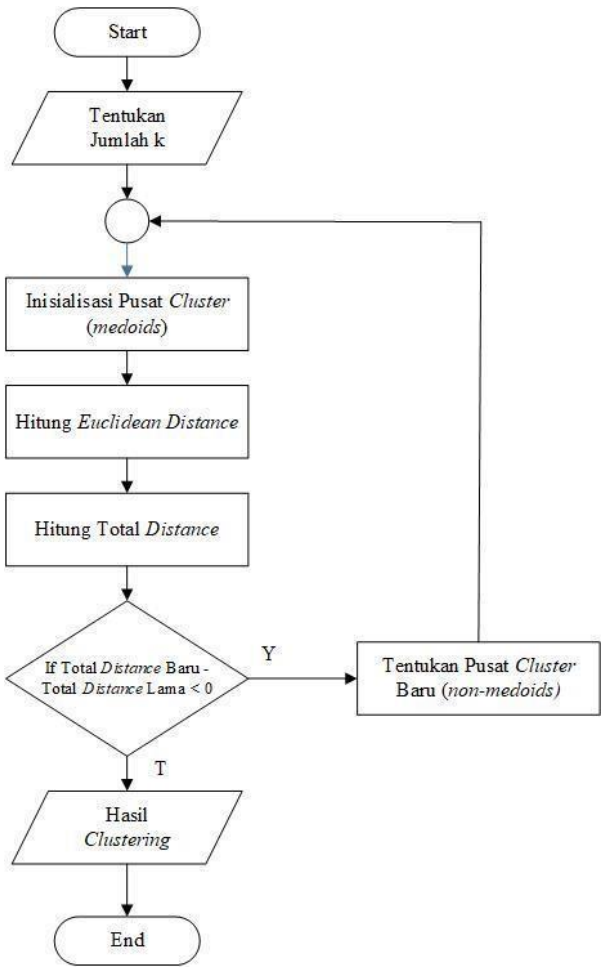
2.6.2 Kekurangan *K-Medoids*

1. Pengelompokannya bekerja dengan baik pada data berskala kecil [20].

2. Kumpulan awal *medoid* yang berbeda dapat menyebabkan pengelompokan akhir yang berbeda. Oleh karena itu, disarankan untuk menjalankan prosedur beberapa kali dengan *set medoid* awal yang berbeda [21].
3. Pengelompokan yang dihasilkan tergantung pada unit pengukuran. Jika *variabel* memiliki sifat yang berbeda atau sangat berbeda dalam hal besarnya, maka disarankan untuk menstandarisasinya [21].

2.6.3 *Flowchart K-Medoids*

Berikut *flowchart* mengenai proses klastering menggunakan algoritma *k-medoids* dapat dilihat pada Gambar 2.6.



Gambar 2.6 *Flowchart* Algoritma *K-Medoids* [49].

2.6.4 Tahapan Perhitungan *K-Medoids*

Berikut tahapan perhitungan dengan menggunakan *k-medoids* [4]:

1. Menentukan jumlah nilai *k* (*cluster* yang akan dibentuk).
2. Menginisialisasi *medoids* pusat *cluster* awal dari masing-masing *cluster* secara acak.
3. Menghitung setiap objek (data observasi) ke *cluster* terdekat menggunakan persamaan ukuran jarak *euclidean distance* dengan rumus persamaan (1).

$$d(x, y) = \sqrt{\Sigma_{i=1}^{n}(xi - yi)^2} \quad (1) \text{ [48]}$$

4. Menghitung secara acak objek masing-masing *cluster* sebagai kandidat *medoid* baru.
5. Menghitung jarak setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru.
6. Menghitung S (total simpangan) 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.

2.7 AHC (*Agglomerative Hierarchical Clustering*)

AHC (*Agglomerative Hierarchical Clustering*) merupakan salah satu teknik dalam mengelompokan data [23]. Teknik ini akan mengelompokan data secara berulang-ulang berdasarkan tingkat kemiripan data satu sama lain dan membentuk sebuah hirarki [24]. Metode hirarki *aglomeratif* memperlakukan setiap titik data sebagai klaster tunggal di awal dan kemudian secara berurutan menggabungkan (atau mengaglomerasi) pasangan klaster hingga semua klaster digabung menjadi satu klaster yang berisi semua titik data [23]. Pada teknik pengelompokan ini menggunakan pendekatan *bottom-up* dengan mengelompokan data mulai dari data yang berupa individu atau *singleton* kemudian digabung menjadi satu kelompok atau *cluster* secara berulang sehingga seluruh data terkelompokan [23]. 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 satu *cluster* terbentuk atau beberapa kondisi yang telah ditentukan terpenuhi [62].

Dalam algoritma AHC (*Agglomerative Hierarchical Clustering*) terdapat tiga teknik pengelompokan [50]:

a. *Single Linkage.*

Teknik menggabungkan objek yang didasarkan pada jarak terkecil.

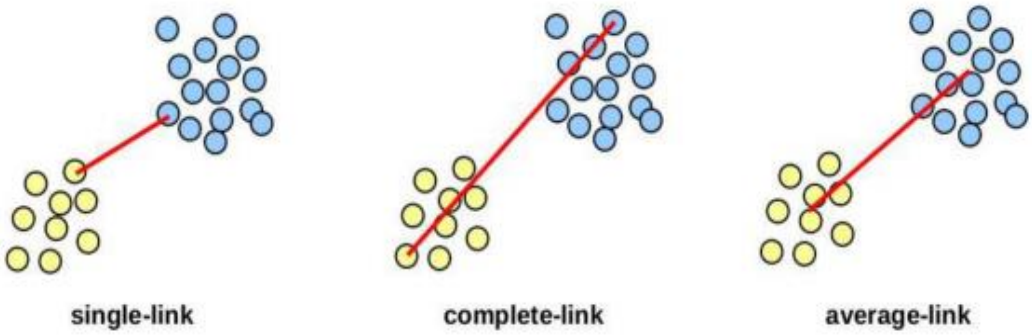
b. *Complete Linkage.*

Teknik menggabungkan objek yang didasarkan pada jarak terjauh.

c. *Average Linkage.*

Teknik menggabungkan objek yang didasarkan pada jarak rata-rata pasangan anggota yang ada dalam himpunan antara dua buah kelompok.

Gambaran *Single Linkage, Complete Linkage* dan *Average Linkage* pada *Agglomerative Hierarchical Clustering* dapat dilihat pada Gambar 2.7.



Gambar 2.7 Perbedaan Teknik Pengelompokan Pada *Agglomerative Hierarchical Clustering* [51].

2.7.1 Kelebihan AHC (*Agglomerative Hierarchical Clustering*)

1. Dapat menghasilkan urutan objek, yang mungkin informatif untuk tampilan data [25].
2. *Cluster* lebih kecil yang dibuat berguna untuk menemukan kesamaan dalam data [25].

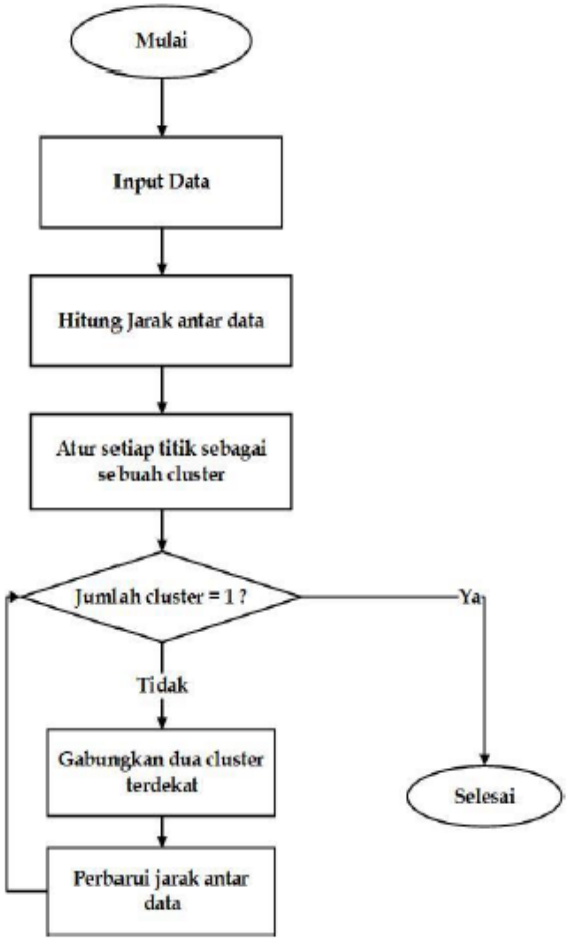
2.7.2 Kekurangan AHC (*Agglomerative Hierarchical Clustering*)

1. Tidak ada ketentuan yang dapat disediakan dalam pendekatan ini untuk relokasi objek yang mungkin telah salah dikelompokkan pada tahap sebelumnya dan hasil yang sama harus diperiksa dengan cermat untuk memastikannya masuk akal [25].
2. Penggunaan berbagai metrik jarak untuk mengukur jarak antar cluster dapat menghasilkan hasil yang berbeda. Oleh karena itu, melakukan beberapa

eksperimen dan kemudian membandingkan hasilnya disarankan untuk membantu kebenaran hasil aslinya [25].

2.7.3 *Flowchart* AHC (*Agglomerative Hierarchical Clustering*)

Berikut *flowchart* mengenai proses klastering menggunakan algoritma AHC (*Agglomerative Hierarchical Clustering*) dapat dilihat pada Gambar 2.8.



Gambar 2.8 *Flowchart* Algoritma AHC (*Agglomerative Hierarchical Clustering*) [50].

2.7.4 Tahapan Perhitungan AHC (*Agglomerative Hierarchical Clustering*)

Berikut tahapan perhitungan dengan menggunakan AHC (*Agglomerative Hierarchical Clustering*) [50].

1. Menghitung jarak menggunakan *euclidean distance*, dengan rumus persamaan (1).

$$x_{rs} = \sqrt{\sum_{k=1}^{p} (y_{rk} - y_{sk})^2} \quad (1)$$

dimana:

x_{rs} : Jarak antara objek r dengan s.

y_{rk} : Nilai objek r pada *variabel* ke-k.

y_{sk} : Nilai objek s pada *variabel* ke-k.

p : Banyaknya *variabel* yang diamati.

2. Penggabungan dua objek terdekat. Apabila nilai jarak dari objek a dan b yang terkecil dibanding jarak antar objek yang lain, maka gabungan dua objek yang pertama yaitu x_ab.
3. Perbarui *matriks* jarak. Apabila x_ab merupakan jarak paling kecil dalam perhitungan jarak yang menggunakan *euclidean distance*, maka rumus untuk metode *agglomerative* dapat dilihat pada persamaan (2), (3) dan (4).

   a. Rumus *Single Linkage.*

   $$d_{(ab)c} = \min\{d_{a,c}; d_{b,c}\} \quad (2)$$

   b. Rumus *Average Linkage*.

   $$d_{(ab)c} = \text{average}\{d_{a,c}; d_{b,c}\} \quad (3)$$

   c. Rumus *Max* atau *Complete Linkage.*

   $$d_{(ab)c} = \max\{d_{a,c}; d_{b,c}\} \quad (4)$$

4. Ulangi kembali langkah 2 dan langkah 3 sampai tersisa satu *cluster*.
5. Membuat *dendogram* [4].

2.8 Penentuan Jumlah *Cluster* Optimal

2.8.1 Metode *Elbow*

Metode *Elbow* merupakan metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan melihat persentase perbandingan jumlah *cluster* yang akan membentuk siku pada suatu titik. Metode ini memberikan ide/gagasan dengan memilih nilai *cluster* kemudian menambahkan nilai *cluster* tersebut untuk dijadikan model data dalam menentukan *cluster* terbaik. Dan selain itu persentase dari perhitungan yang dihasilkan merupakan perbandingan antara jumlah *cluster* yang ditambahkan [52].

Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan

nilai *cluster* kedua memberikan sudut pada grafik nilai penurunan terbesar maka nilai *cluster* tersebut adalah yang terbaik. Untuk mendapatkan perbandingan adalah menghitung SSE (*Sum of Square Error*) dari setiap nilai *cluster*. Karena semakin besar jumlah *cluster k* maka nilai SSE akan semakin kecil.[52]

Berikut tahapan metode *Elbow* dalam menentukan nilai *k* pada algoritma *k-means* [52]:

1. Inisialisasi nilai awal *k*.
2. Menaikkan nilai *k*.
3. Menghitung hasil *sum of square error* dari setiap nilai *k*.
4. Analisis hasil *sum of square error* dari nilai *k* yang mengalami penurunan drastis.
5. Temukan dan atur nilai *k* berbentuk siku.

Pada metode *elbow* nilai *cluster* terbaik akan diambil dari nilai *Sum of Square Erro*r (SSE) yang memiliki penurunan signifikan dan berbentuk siku.

2.8.2 *Silhouette Method*

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation*. Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut [13]:

1. Hitung rata-rata jarak dari suatu dokumen misalkan *i* dengan semua dokumen lain yang berada dalam satu *cluster*.
   $$a(i) = \frac{1}{|A|-1} \sum j \in_{A, j \neq i} d(i,j) \quad (1)$$
   dengan *j* adalah dokumen lain dalam satu *cluster A* dan *d(i,j)* adalah jarak antara dokumen *i* dengan *j*.
2. Hitung rata-rata jarak dari dokumen *i* tersebut dengan semua dokumen di *cluster* lain, dan diambil nilai terkecilnya.
   $$d(i,C) = \frac{1}{|A|} \sum j \in C \; d(i,j) \quad (2)$$
   dengan *d(i,C)* adalah jarak rata-rata dokumen *i* dengan semua objek pada *cluster* lain *C* dimana $A \neq C$.
   $$b(i) = \min C \; \neq A \, d(i,C) \quad (3)$$
3. Nilai *Silhouette Coefficient* nya adalah :
   $$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

(4)

2.9 Evaluasi Model

Evaluasi adalah proses pemberian makna atau ketetapan kualitas hasil pengukuran dengan cara membandingkan angka hasil pengukuran tersebut dengan kriteria tertentu. Kriteria sebagai pembanding dari proses pengukuran atau dapat pula ditetapkan sesudah pelaksanaan pengukuran. Kriteria ini dapat berupa proses/kemampuan rata-rata unjuk kerja kelompok dan berbagai patokan yang lain. Evaluasi merupakan proses yang sistematis dan berkelanjutan untuk mengumpulkan, mendeskripsikan, menginterpretasikan, dan menyajikan informasi tentang suatu program untuk dapat digunakan sebagai dasar membuat keputusan, menyusun kebijakan maupun menyusun program selanjutnya. Adapun tujuan evaluasi adalah untuk memperoleh informasi yang akurat dan objektif tentang suatu program. Informasi tersebut dapat berupa proses pelaksanaan program, dampak/hasil yang dicapai, efisiensi serta pemanfaatan hasil evaluasi yang difokuskan untuk program itu sendiri, yaitu untuk mengambil keputusan apakah dilanjutkan, diperbaiki atau dihentikan. Selain itu, juga dipergunakan untuk kepentingan penyusunan program berikutnya maupun penyusunan kebijakan yang terkait dengan program [53].

Model evaluasi merupakan desain evaluasi yang dikembangkan oleh para ahli evaluasi, yang biasanya dinamakan sama dengan pembuatnya atau tahap evaluasinya [54].

2.9.1 *Average Within* dan *Average Between*

Untuk mengetahui metode mana yang mempunyai kinerja terbaik, dapat digunakan rata-rata simpangan baku baku dalam *cluster* (S_W) dan simpangan baku antar *cluster* (S_B) [55].

Rumus rata-rata simpangan baku dalam *cluster* (S_W) [54]:

$$S_W = K^{-1} \sum_{k=1}^{K} S_k \quad (1)$$

Dimana :

K = banyaknya *cluster* yang terbentuk.

S_k = simpangan baku *cluster* ke-k.

Rumus simpangan baku antar *cluster* (SB):

$$S_B = [(K-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{X})^2]^{1/2} \quad (2)$$

Dimana:

$\bar{X}_k$ = rataan *cluster* ke-k.

$\bar{X}$ = rataan keseluruhan *cluster*.

Metode yang mempunyai rasio terkecil merupakan metode terbaik. *Cluster* yang baik adalah *cluster* yang mempunyai *homogenitas* (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*) dan *heterogenitas* yang tinggi antar *cluster* yang satu dengan *cluster* yang lain (*between cluster*) [55].

2.9.2 *Davies Bouldin Index* (DBI)

Davies Bouldin Index (DBI) merupakan metode untuk mengecek hasil *clustering*. Pendekatan pengujian nilai DBI berupa nilai separasi dan kohesi. Kohesi berupa jumlah dari kemiripan data terhadap pusat *cluster* dari *cluster* tersebut. Separasi adalah jarak antara pusat *cluster* dari *cluster*. *Cluster* yang optimal adalah *cluster* yang memilki nilai separasi yang tinggi dan nilai kohesi yang rendah. Nilai *Davies Bouldin Index (DBI)* yang semakin mendekati nilai 0 menandakan semakin baik *cluster* yang diperoleh. Semakin rendah nilai DBI menunjukkan hasil *cluster* yang optimal. [56]

Sum of square within cluster (SSW) adalah Persamaan untuk mengetahui *matrik* kohesi dalam sebuah *cluster* ke-*i* yang dapat dilihat pada Persamaan 1.

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad (1)$$

Keterangan :

mi = jumlah data dalam *cluster* ke-*i*

ci = *centroid cluster* ke-*i*

d (xj, ci) = *jarak euclidean* setiap data ke *centroid*

Sum of square between cluster (SSB) adalah persamaan untuk mengetahui nilai separasi antara *cluster* yang dapat dilihat pada Persamaan 2.

$$SSB_{i,j} = d(c_i, c_j) \quad (2)$$

Keterangan:

d (c_i, c_j) = jarak antar *centroid*

Setelah nilai separasi dan kohesi diperoleh, lalu dilakukan pengukuran rasio (R_{ij}) untuk mengetahui nilai perbandingan antara *cluster* ke-*i* dan *cluster* ke-*j*, sesuai dengan persamaan 3.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{3}$$

Persamaan untuk menghitung nilai *Davies Bouldin Index* (DBI) sesuai dengan persamaan 4.

$$DBI = \frac{1}{K}\sum_{i=1}^{k} max_{i \neq j}\,(R_{i,j}) \tag{4}$$

Keterangan:

k = jumlah *cluster* yang digunakan.

Semakin rendah nilai *Davies Bouldin Index* (DBI) yang diperoleh, maka semakin baik kualitas *cluster* yang diperoleh dari suatu *clustering* data [56].

2.9.3 *Silhouette Index*

Silhouette Index adalah teknik validasi untuk mengukur seberapa dekat setiap objek dalam satu *cluster* dan objek di *cluster* lain serta mengetahui seberapa baik objek terletak dalam *cluster*nya. Nilai *Silhouette Index* berada di antara -1 hingga +1 dan hasil nilai maksimum menunjukan jumlah *cluster* yang optimal [48].

2.9.4 Rasio Simpangan Baku

Pemilihan metode yang menghasilkan kualitas pengelompokan terbaik dilakukan dengan memperhatikan nilai rasio rata-rata simpangan baku dalam *cluster* dan simpangan baku antar *cluster*. Rata-rata simpangan baku di dalam *cluster* (Sw) dinyatakan dengan:

$$S_w = \frac{1}{c}\sum_{k=1}^{c} S_k \tag{1}$$

Simpangan baku antar *cluster* (S_b) dinyatakan sebagai

$$S_b = \left[\frac{1}{c-1}\sum_{k=1}^{c}\left(\overline{X}_k - \overline{X}\right)^2\right]^{\frac{1}{2}} \tag{2}$$

c adalah jumlah *cluster*, S_k merupakan simpangan baku di dalam *cluster* ke-k. $\bar{X}_k$ sebagai rata-rata *cluster* ke-k dan $\bar{X}$ adalah rata-rata dari semua *cluster*. Semakin kecil nilai S_w dan semakin besar nilai S_b maka metode tersebut memiliki kinerja yang baik, artinya memiliki *homogenitas* yang tinggi. Metode yang dipilih adalah yang memberikan nilai S_w/S_b terkecil [57].

2.10 Bahasa Pemrograman R

2.10.1 Sejarah R

R Merupakan bahasa yang digunakan dalam komputasi statistik yang pertama kali dikembangkan oleh Ross Ihaka dan Robert Gentlement di University of Auckland New Zealand yang merupakan akronim dari nama depan kedua pembuatnya. Sebelum R dikenal ada S yang dikembangkan oleh John Chambers dan rekan-rekan dari Bell Laboratories yang memiliki fungsi yang sama untuk komputasi statistik. Hal yang membedakan antara keduanya adalah *R* merupakan sistem komputasi yang bersifat gratis [58]. Logo *R* dapat dilihat pada Gambar 2.9.

Gambar 2.9 Logo *R*

R dapat dibilang merupakan aplikasi sistem statistik yang kaya. Hal ini disebabkan banyak sekali Paket yang dikembangkan oleh pengembang dan komunitas untuk keperluan analisa statistik seperti *linear regression*, *clustering*, *statistical test*, dll. Selain itu, *R* juga dapat ditambahkan Paket-Paket lain yang dapat meningkatkan fiturnya. Sebagai sebuah bahasa pemrograman yang banyak digunakan untuk keperluan analisa data, *R* dapat dioperasikan pada berbagai sistem operasi pada komputer. Adapun sistem operasi yang didukung antara lain: UNIX, Linux, *Windows*, dan MacOS [58].

2.10.2 Fitur dan Karakteristik R

R memiliki karakteristik yang berbeda dengan bahasa pemrograman lain seperti C++, python, dll. R memiliki aturan/fungsi yang berbeda dengan bahasa pemrograman yang lain yang membuatnya memiliki ciri khas tersendiri dibanding bahasa pemrograman yang lain [58]. Beberapa ciri dan fitur pada R antara lain [58]:

1. **Bahasa R bersifat case sensitif**.

   Maksudnya adalah dalam proses *input* R huruf besar dan kecil sangat diperhatikan.
2. **Segala sesuatu yang ada pada program R akan diangap sebagai objek**.

   Konsep objek ini sama dengan bahasa pemrograman berbasis objek yang lain seperti Java, C++, python, dll.Perbedaannya adalah bahasa R relatif lebih sederhana dibandingkan bahasa pemrograman berbasis objek yang lain.
3. ***Interpreted language* atau *script***.

   Bahasa R memungkinkan pengguna untuk melakukan kerja pada R tanpa perlu kompilasi kode program menjadi bahasa mesin.
4. Mendukung proses *loop*, *decision making*, dan menyediakan berbagai jenis *operator* (aritmatika, logika, dll).
5. Mendukung *export* dan *import* berbagai format *file*, seperti:TXT, CSV, XLS, dll.
6. **Mudah ditingkatkan melalui penambahan fungsi atau *library***.

   Penambahan Paket dapat dilakukan secara online melalui *CRAN* atau melalui sumber seperti *github*.
7. **Menyedikan berbagai fungsi untuk keperluan visualisasi data**. *Visualisasi* data pada R dapat menggunakan paket bawaan atau paket lain seperti *ggplot2*, *ggvis*, dll.

2.10.3 Kelebihan dan Kekurangan R

Selain karena R dapat digunakan secara gratis terdapat kelebihan lain yang ditawarkan, antara lain [58]:

1. ***Protability***.

   Penggunaan *software* dapat digunakan kapanpun tanpa terikat oleh masa berakhirnya lisensi.

2. ***Multiplatform***.

   R bersifat *Multiplatform Operating Systems*, dimana *software* R lebih kompatibel dibanding *software* statistika lainnya. Hal ini berdampak pada kemudahan dalam penyesuaian jika pengguna harus berpindah sistem operasi karena R baik pada sistem operasi seperti *windows* akan sama pengoperasiannya dengan yang ada di Linux (paket yang digunakan sama).

3. ***General*** dan ***Cutting-edge***.

   Berbagai metode statistik baik metode klasik maupun baru telah diprogram kedalam R. Dengan demikian *software* ini dapat digunakan untuk analisis statistika dengan pendekatan klasik dan pendekatan modern.

4. ***Programable***.

   Pengguna dapat memprogram metode baru atau mengembangkan modifikasi dari analisis statistika yang telah ada pada sistem R.

5. **Berbasis analisis matriks**.

   Bahasa R sangat baik digunakan untuk *programming* dengan basis *matriks*.

6. Fasilitas grafik yang lengkap.

Adapun kekurangan dari R antara lain [58]:

1. ***Point and Click GUI***.

   Interaksi utama dengan R bersifat CLI (*Command Line Interface*), walaupun saat ini telah dikembangkan Paket yang memungkinkan kita berinteraksi dengan R menggunakan GUI (*Graphical User Interface*) sederhana menggunakan Paket *R-Commander* yang memiliki fungsi yang terbatas. *R-Commander* sendiri merupakan *GUI* yang diciptakan dengan tujuan untuk keperluan pengajaran sehingga analisis statistik yang disediakan adalah yang klasik. Meskipun terbatas Paket ini berguna jika kita membutuhkan analisis statistik sederhana dengan cara yang simpel.

2. ***Missing statistical function***.

   Meskipun analisis statistika dalam R sudah cukup lengkap, namun tidak semua metode statistika telah diimplementasikan ke dalam R. Namun karena R merupakan *lingua franca* untuk keperluan komputasi statistika modern staan ini, dapat dikatakan ketersediaan fungsi tambahan dalam bentuk Paket hanya masalah waktu saja.

2.11 *Microsoft Excel*

Microsoft excel atau *microsoft office excel* atau *excel* adalah sebuah program aplikasi lembar kerja *spreadsheet* yang dibuat dan didistribusikan oleh Microsoft Corporation untuk *system* operasi *Microsoft Windows* dan Mac OS. Aplikasi ini memiliki fitur kalkulasi dan pembuatan grafik dengan menggunakan strategi *marketing Microsoft* yang agresif, menjadikan *microsoft excel* sebagai salah satu program komputer yang populer digunakan di dalam *computer* mikro hingga saat ini. Bahkan, saat ini program *microsoft excel* merupakan program *spreadsheet* paling banyak digunakan oleh banyak pihak, baik di *platform* PC berbasis *Windows* maupun *platform* Macintosh berbasis Mac OS [59].

2.12 *RStudio*

RStudio adalah *Integrated Development Environment* (IDE) baru untuk bahasa pemrograman *R*. RStudio adalah proyek sumber terbuka yang dimaksudkan untuk menggabungkan berbagai komponen *R* (konsol, pengeditan sumber, grafik, riwayat, bantuan, dan lain-lain.) menjadi terintegrasi dan produktif. *RStudio* dirancang untuk memudahkan pembelajaran mengenai kurva/grafik bagi pengguna *R* baru serta menyediakan alat produktivitas tinggi untuk pengguna yang lebih mahir. *RStudio* berjalan di semua *platform* utama termasuk *Windows*, Mac OS X, dan Linux. Selain aplikasi *desktop*, RStudio dapat digunakan sebagai *server* untuk mengaktifkan akses *web* ke sesi *R* yang berjalan pada sistem jarak jauh [60].

2.13 Penelitian Terdahulu

Pada tesis ini, penulis membandingkan beberapa penelitian terdahulu mengenai segmentasi pasar berdasarkan penjualan produk menggunakan metode *clustering* baik *k-means, k-medoids* maupun ahc (*agglomerative hierarchical clustering*) yang dapat dilihat pada Tabel 2.1.

Tabel 2.1 Penelitian Terdahulu

No	Nama Penulis	Judul	Hasil Penelitian	Persamaan	Perbedaan	
					Penelitian Terdahulu	Penelitian Ini
1	Sabbir Hossain Shihab, Shyla Afroge and Sadia Zaman Mishu [56]	*RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A*	Penelitian ini menunjukkan bahwa versi lanjutan *k-means* (*advanced k-means*) secara efektif dapat mengurangi total waktu masing-masing sebesar 27,8% dan 97,8% dibandingkan dengan *agglomerative* dan *agglomerative* menghasilkan pengelompokan yang lebih baik	Segmentasi pasar yang dilakukan menggunakan lanjutan *k-means* (*advanced k-means*) dan AHC (*Agglomerative Clustering*)	Membandingkan 2 metode yaitu lanjutan *k-means* (*advanced k-means*) dan *Agglomerative Clustering* pada segmentasi pasar berdasarkan RFM.	Adanya tambahan metode yang dibandingkan yaitu *K-Medoids* pada analisis segmentasi pasar menggunakan *software* RStudio.

		Comparative Study	daripada *k-means* standar sehubungan dengan jarak intra cluster dan jarak antar cluster.			
2	Phan Duy Hung, Nguyen Duc Ngoc and Tran Duc Hanh [57]	*K-Means Clustering Using R A Case Study of Market Segmentation*	Penelitian ini didapat jumlah *centroid* optimal adalah 4 dalam dua dimensi atribut: Usia dan Pembelian dari kumpulan data *Black Friday*. Dengan begitu, pengecer dan penyedia dapat memanfaatkan hasil ini untuk meluncurkan strategi pemasaran yang tepat dan melihat hasil yang lebih baik.	Segmentasi pasar dengan mengunakan metode *K-means Clustering* menggunakan Bahasa Pemrograman R	Hanya menggunakan 1 metode yaitu *K-Means*.	Tidak hanya menggunakan 1 metode tetapi menggunakan 3 metode yaitu *K-Means, K-Medoids* dan AHC (*Agglomerative Hierarchical Clustering*).
3	Geraldi Catur Pamuji dan H Rongtao [39]	*A Comparison study of DBScan and K-Means Clustering in* Jakarta *rainfall based on the Tropical Rainfall*	Pada penelitian ini setiap algoritma membentuk jumlah *cluster* yang berbeda. *K-Means* menggunakan variabel yang telah ditentukan (k) dengan nilai 3, dan *DBScan* memiliki kemampuan untuk menentukan berapa banyak *cluster*, yang	Penerapan menggunakan metode *K-Means Clustering*.	Membandingkan *DBScan* dan *K-Means Clustering* pada curah hujan Jakarta berdasarkan Curah Hujan Tropis.	Adanya tambahan metode yang dibandingkan yaitu *K-Medoids* dan AHC (*Agglomerative Hierarchical Clustering*) pada analisis segmentasi pasar berdasarkan penjualan produk.

		Measuring Mission (TRMM) 1998-2007	dapat dibentuk berdasarkan titik data. *K-means* lebih cepat dalam memproses kumpulan data yang lebih besar. *K-Means* menghasilkan hasil yang lebih efisien dan akurat daripada *DBScan*.			
4	Juniar Hutagalung, Muhammad Syahril dan Sobirin [58]	*Implementation of K-Medoids Clustering Method for Indihome Service Package Market Segmentation*	Penelitian segmentasi pasar menggunakan metode *K-medoids* sangat berguna untuk memahami preferensi dan perilaku konsumen, meningkatkan kepuasan pelanggan, dan merancang strategi pemasaran yang lebih efektif. Sistem yang telah dirancang dapat digunakan sebagai solusi permasalahan secara tepat dan akurat. Pengolahan data paket layanan Indihome menggunakan metode clustering *k-medoids* berupa	Implementasi metode *K-Medoids* pada segmentasi pasar.	Menggunakan 1 metode yaitu *K-Medoids* untuk segmentasi pasar paket layanan indihome.	Menggunakan 3 metode yatu *K-Means, K-Medoids* dan AHC (*Agglomerative Hierarchical Clustering*) untuk analisis segmentasi pasar berdasarkan penjualan produk.

			anggota cluster STO (*Sentral Telephone Automated*) yang sangat potensial, potensial, dan tidak potensial.			
5	Arief Fathoni Argadian [59].	Analisis Segmentasi Pasar Produk Kopi Celup di Kota Yogyakarta, Kabupaten Sleman, dan Bantul.	Penelitian ini bertujuan untuk mengidentifikasi dan menentukan segmen pasar kopi celup dengan metode *k-means* dan deskriptif di Kota Yogyakarta, Kabupaten Sleman, dan Bantul dimana konsumen kopi celup dibagi menjadi tiga segmen kemudian dianalisis untuk menentukan segmen pasar kopi celup unggulan berdasarkan demografi dan *psikografis*. Berdasarkan penelitian ini, hasil identifikasi menunjukan bahwa terdapat tiga segmen pasar kopi celup yaitu segmen konsumen pengikut keluarga, konsumen	Analisis segmentasi pasar menggunakan Metode *K-Means.*	Menggunakan 1 metode yaitu *K-Means* untuk analisis segmentasi pasar produk kopi celup.	Adanya tambahan metode yang digunakan yaitu *K-Medoids* dan AHC (*Agglomerative Hierarchical Clustering*) untuk analisis segmentasi pasar berdasarkan penjualan produk.

			pengikut tren dan konsumen yang sibuk dengan aktivitas sehari-hari. Segmen yang sibuk dengan aktivitas sehari-hari merupakan segmen unggulan karena memiliki anggota yang lebih banyak dibandingkan segmen lainnya. Segmen ini juga lebih sering mengkonsumsi dan membeli kopi celup.			
6	Putri Eka Prakasawati, Yulison H. Chrisnanto, dan Asep Id Hadiana [60]	Segmentasi Pelanggan Berdasarkan Produk Menggunakan Metode *K-Medoids*	Dalam penelitian ini sebuah sistem mengelompokkan segmentasi pelanggan berdasarkan produk. Proses pada sistem pengelompokan pelanggan ini menggunakan sebuah algoritma *K-Medoid clustering* untuk mengelompokkan pelanggan berdasarakan segmentasi pada produk jumlah pembelian dan	Menggunakan metode *K-Medoids Clustering*.	Segmentasi pelanggan menggunakan metode *K-Medoids Clustering*	Analisis segmentasi pasar menggunakan 3 metode yaitu *K-Means*, *K-Medoids* dan AHC (*Agglomerative Hierarchical Clustering*).

			area. Dengan data uji sebanyak 6 daerah, 7000 data outlet dan 28 produk. Dapat disimpulkan menggunakan 3 cluster berikut hasilnya C1 menghasilkan 27%, C2 menghasilkan 43% dan C3 menghasilkan 30%.			
7	Sundary Ayu Pratiwi, Muhammad Syahril, dan Rini Kustini [63]	Segmentasi Pasar Penjualan Unit Mobil Menggunakan Metode *Agglomerative Hierarchical Clustering (AHC)*	Pada penelitian ini sistem mampu memecahkan permasalahan di dalam menemukan segmentasi pasar penjualan dan dapat diterapkan sebagai bahan pengambilan keputusan dalam meningkatkan strategi pemasaran. Penerapan metode *agglomerative hierarchical clustering* dilakukan dengan cara menghitung nilai rata-rata setiap *variable*, menghitung nilai standar deviasi, menghitung nilai *zero standard*,	Segmentasi pasar menggunakan Metode *Agglomerative Hierarchical Clustering* (AHC)	Penerapan 1 metode yaitu *Agglomerative Hierarchical Clustering* (AHC) pada segmentasi pasar penjualan unit mobil.	Penerapan 3 metode yaitu *K-Means, K-Medoids dan Agglomerative Hierarchical Clustering* pada analisis segmentasi pasar berdasarkan penjualan produk.

			menghitung nilai pengukuran jarak dan melakukan pengelompokkan menggunakan *euclidean single linkage*.			