BAB II
KAJIAN PUSTAKA

2.1. **Penelitian Relevan**

Berikut merupakan penelitian – penelitian terdahulu yang relevan dengan penelitian yang akan dilakukan :

Penelitian pertama dilakukan oleh (Rohalidyawati, Rahmawati, & Mustafid, 2020) dalam jurnal yang berjudul "Segmentasi Pelanggan E-Money Dengan Menggunakan Algoritma Dbscan (Density Based Spatial Clustering Applications With Noise) di Provinsi Dki Jakarta". Dalam penelitian tersebut dijelaskan bahwa Dalam menentukan target pemasaran pada pelanggan E-money diperlukan cara yang efektif yaitu dengan membuat segmen – segmen pasar,untuk mempermudah dalam menganalisis cluster. Dimana dalam penelitian ini dilakukan metode clustering yang dapat digunakan adalah DBSCAN (*Density Based Spatial Clustering Applications with Noise).*Algoritma DBSCAN termasuk algoritma nonparametrik dalam *unsupervised learning* . untuk membentuk pelanggan yang potensial serta menetukan kualitas dari segmen – segmen yang telah terbentuk. Hasil dari pengujian ini bertujuan menerapkan algoritma DBSCAN (*Density Based Spatial Clustering Application with Noise*) dalam membentuk segmen-segmen dari pelanggan E-money di provinsi DKI Jakarta untuk menentukan pelanggan potensial serta menentukan kualitas dari segmen-segmen yang telah terbentuk dengan menggunakan *Silhouette Coefficient.*

Penelitian Kedua dilakukan oleh (Sulistyowati, Ketherin, Arifiyanti, & Sodik, 2018) yang berjudul "Analisis segmentasi Konsumen menggunakan Algorima K-Means Clustering". Dalam penelitian tersebut dijelaskan bahwa dalam menentukan segmentasi konsumen berdasarkan data pembelian sepeda motor yang sebelumnya proses segmnetasi konsumen masih dilakukan secara manual sehinggan menyebabkan kesulitan dalam pengelompokkan data dan sering terjadi kesalahan dalam pengelompokkan. Melalui penelitian ini, dilakukan sebuah analisa terhadap karakteristik konsumen di dealer Honda agar dapat menentukan segmentasi konsumen dengan menerapkan metode K-Means. Tujuan penelitian ini adalah untuk menghasilkan segmentasi konsumen yang dapat dimanfaatkan

divisi marketing,untuk memilih strategi promosi yang sesuai dengan karakteristik konsumen, sehingga promosi berjalan dengan efektif dan efisien. Penelitian ini menggunakan Algoritma K-Means Clustering. Pengumpulan data penelitian ini menggunakan metode observasi, wawancara, dan analisi. Hasil dari perbandingan aplikasi adalah selisih dari masing- masing cluster dengan jumlah rata – rata 7%.

Penelitian ketiga yang dilakukan oleh (Harani, Prianto, & Nugraha, 2020) yang berjudul "Segmentasi Pelanggan Produk Service Digital Indihome Menggunakan Algoritma K-Means berbasis Python" dalam penelitian ini dilakukan untuk analisa segmentasi karakteristik pelanggan sebagai dasar penetapan segmentasi pelanggan untuk mengetahui perilaku pelanggan dan menerapkan strategi pemasaran yang tepat Metode yang digunakan dalam penelitian ini adalah Studi Pustaka dan pengumpulan data. Hasil dari penelitian ini menunjukkan bahwa simulasi terbaik berdasarkan evaluasi cluster dengan presentasi data train 50% dan data test 50%. Dimana masing – masing cluster dari ketiga simulasi tersebut adalah cluster 0 memiliki anggota 396 pelanggan dengan kategori pelanggan yang memberikan keuntungan terbesar bagi perusahaan, cluster 1 memiliki anggota 286 pelanggan dengan kategori pelanggan yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan, dan cluster 2 memiliki anggota14 pelanggan dengan kategori pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan.

Dari ketiga jurnal diatas kita dapat menyimpulkan bahwa perbedaan terletak pada metode Clustering yang digunakan dalam penelitian pertama menggunakan metode Dbscan (Density Based Spatial Clustering Applications With Noise) yang berfokus pada dalam menganalisis cluster dengan membentuk dan menentukan kualitas dari segmen-segmen yang telah terbentuk dengan menggunakan *Silhouette Coefficient.*

Perbedaan yang ada dalam penelitian Kedua dan ketiga yaitu metode clustering yang digunakan adalah K-Means Clustering di penelitian kedua ini dijelaskan bahwa dalam menentukan segmentasi konsumen dilakukan sebuah analisa terhadap karakteristik konsumen sehingga menghasilkan segmentasi konsumen dan strategi promosi yang sesuai dengan karakteristik konsumen dan

penelitian ketiga perbedaan terdapat dalam menentukan segmentasi karakteristik pelanggan untuk mengetahui perilaku pelanggan dan menghasilkan simulasi terbaik melalui evaluasi cluster dengan presentasi data train. Dan dapat disimpulkan bahwa perbedaan dari ketiga jurnal dengan penelitian ini adalah dari masing – masing metode clustering yang digunakan berbeda dan perbedaan dalam proses pengujian data yang diterapkan.

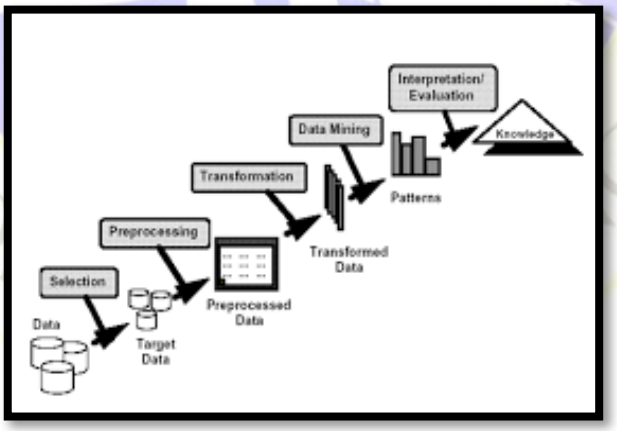
1.2. Landasan Teori

Pada landasan teori ini peneliti akan membahas tentang pengertian *Data Mining, Segmentasi Pelanggan, Clustering, Algomerative hierarchical Clustering.*

2.2.1. Data Mining

Data mining adalah proses yang menggunakan teknik statistic, matematika kecerdasan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.

Istilah data mining dan knowledge discovery in databases (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining,* proses KDD *(knowledge discovery in databases)* secara garis besar dapat dijelaskan sebagai berikut (Prasetya, 2018)



Gambar 2. 1 *Tahapan* Knowledge Discovery in Databases

Sumber : https://www.teorikomputer.com

1. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam knowledge data discovery (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Preprocessing atau Cleaning

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus knowledge data discovery. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak juga dilakukan proses enrichment, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi.

3. Transformation

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam knowledge data discovery merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat tergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation atau evaluation

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada pada sebelumnya.

2.2.2. **Segmentasi**

Segmentasi terus menjadi konsep pemasaran yang penting juga dalam konteks relationship marketing. Meningkatkan hubungan dengan pelanggan menjadi lebih menarik dan akan

menghasilkan pemahaman yang lebih baik tentang kebutuhan pelanggan Menurut (Sahalin, 2019) Segmentasi adalah proses membagi pelanggan menjadi beberapa cluster dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran. Segmentasi pelanggan adalah salah satu langkah awal dalam membuat model bisnis.

2.2.3. **Clustering**

Clustering merupakan proses mengelompokkan atau penggolongn objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan atar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas / cluster. Clustering bertujuan untuk menemukan kelompok (cluster) objek yang berguna, dimana penggunaannya tergantung dari tujuan analisa data.

Ada dua jenis tipe clustering, yaitu *patitional clustering* dan *hierarchical clustering*, dengan *patitional clustering*, objek data dibagi ke dalam sub-himpunan (cluster) yang tidak overlap sedemikian hingga tiap objek data berada dalam tepat satu sub-himpunan. K-Means merupakan algoritma yang menggunakan *patitional clustering*, yang menggabungkan semua objek atau membuat pengelompokkan data. *hierarchical clustering* merupakan sebuah cluster bersarang sayng diatur sebagai pohon hirarki. Tiap simpul (cluster) dalam pohon mrupakan gabungan dari anaknya (*subcluster*) dan simpul akar berisi semua objek (Fajar & Rahayu, 2018).

Analisis cluster merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis cluster mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam cluster yang sama. Solusi analisis cluster bersifat tidak unik, anggota cluster untuk tiap penyelesaian/solusi tergantung pada beberapa elemen prosedur dan beberapa solusi yang berbeda dapat diperoleh dengan mengubah satu elemen atau lebih. Solusi cluster secara keseluruhan bergantung pada variabel-variaabel yang digunakan sebagai dasar untuk menilai kesamaan. Penambahan atau pengurangan variabel-variabel yang relevan dapat mempengaruhi substansi hasi analisis cluster.

2.2.4. **Agglomerative Hierarchical Clustering (AHC)**

Strategi pengelompokkan *hierarchical clustering* umumnya ada dua jenis yaitu *Agglomerative* (*Bottom-up*) dan *Devisive* (*Top-Down*). Namun dalam penelitian ini

menggunakan pendekatan *agglomerative hierarchical clustering*. Dalam algoritma *agglomerative clustering* dihasilkan pengelompokan data yang dapat dilihat dengan dendrogam tidak diperlukan penentuan jumlah *cluster* pada awal pengelompokkan, dan *agglomerative hierarchical clustering* dengan pendekatan bawah-atas (*bottom-up*) dimana pengelompokkan data dimulai dari kecil ke pengelompokkan yang besar.

Agglomerative hierarchical clustering (AHC) dengan menggunakan *bottom-up*, dimulai dari masing-masing data sebagai sebuah *cluster*, kemudian akan digabungkan menjadi kelompok yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak keatas membentuk hirarki (Sahalin, 2019),.

Dalam *agglomerative* terbagi menjadi 3 teknik pengelompokkan kedekatan jarak bersifat *bottom-up* yaitu *Single Linkage* (jarak terdekat), *Average Linkage* (jarak rata – rata), dan *Complete Linkage* (jarak terjauh).

Membentuk Matrik jarak, Misal dengan menggunakan Manhattan Distance atau Euclidian Distance.

- *Manhattan Distance* :

$$D = \sum_{i=1}^{n} ¿ bi - ai \vee ¿ ¿ \quad \text{.....(2.1)}$$

- *Euclidian Distance* :

$$D\ (a,b) = \sqrt{\sum_{i=1}^{n} (bi - ai)^2}. \quad \text{.....(2.2)}$$

Terdapat 3 metode pengelompokkan *Agglomerative Hierarchical Clustering*, Berikut 3 metode tersebut yaitu:

a. **Single Linkeage** (Jarak Terdekat)

Menentukan kedekatan kelompok diantara 2 kelompok terdekat (terkecil) antara 2 pengelompokkan data yang berbeda. Formulasi *single linkage* adalah:

$$d_{uv} = min\ \{d_{uv}\},\ d_{uv} \in D \quad \text{.....(2.3)}$$

Keterangan: $\{d_{uv}\}$ adalah jarak antara 2 data yaitu U dan V pada masing-masing *cluster* U dan V.

b. **Complete Linkage** (Jarak Terjauh)

Menentukan kelompok kedekatan diantara 2 kelompok terjauh (terbesar) antara 2 data pengelompokkan data yang berbeda. Formulasi untuk *complete linkage* adalah:

$d_{uv} = max\ \{d_{uv}\},\ d_{uv} \in D$ .....(2.4)

Keterangan : $\{d_{uv}\}$ adalah jarak antara 2 data yaitu U dan V dari masing-masing *cluster* U dan V.

c. **Average Linkage** (Jarak Rata – rata)

Menentukan kedekatan diantara dua kelompok dari jarak rata-rata antar dua data dari *cluster* yang berbeda. Formulasi untuk *average linkage* adalah :

$d_{uv} = average\ \{d_{uv}\},\ d_{uv} \in D$ .....(2.5)

keterangan : |U| dan |V| adalah jumlah data yang ada dalam *cluster* U dan V.

Dengan menggunakan formulasi *single linkage, average linkage, dan complete linkage* akan menghasilkan dendogram.

Langkah – langkah Algoritma *Agglomerative Hierarchical Clustering* dapat dijabarkan sebagai berikut (Prasetya, 2018):

1. Hitung matrik kedekatan berdasarkan jenis jarak yang digunakan.
2. Ulangi langkah 3 dan 4, hingga hanya satu *cluster* yang tersisa.
3. Gabungkan dua *cluster* terdekat berdasarkan parameter kedekatan yang ditentukan.
4. Perbarui matriks kedekatan untuk merefleksi kedekatan diantara *cluster* baru dan *cluster* asli yang sudah digabungkan