BAB II
TINJAUAN PUSTAKA

II.1. Tinjauan Tempat Penelitian

Untuk medukung pembuatan laporan penelitian, maka perlu dijelaskan beberapa hal mengenai tempat penelitian. Tinjauan tempat penelitian meliputi profil perusahaan, logo perusahaan, serta struktur organisasi dan deskripsi kerja.

II.1.1. Profil Perusahaan

Jeger Jersey Indonesia adalah *startup* yang bergerak dalam bidang penjualan *merchandise* tim sepakbola Indonesia. *Startup* ini didirikan pada tahun 2011 di kota Bandung. Jeger Jersey Indonesia fokus pada penjualan produk melalui media *online.* Sempat beberapa kali melakukan perubahan nama untuk toko *online* yang dikembangkan*,* akhirnya dipilih *brand* dengan nama Jeger Jersey. Toko *online* ini menjual berbagai produk seperti *jersey*, *polo shirt*, tas, jaket dan lainnya. Setelah empat tahun berdiri, Jeger Jersey kini semakin dikenal di kalangan pencinta sepakbola Indonesia.

II.1.2. Logo Perusahaan

Gambar II.1 merupakan logo dari Jeger Jersey Indonesia. Logo berikut juga digunakan sebagai *brand* dari toko *online* Jeger Jersey.



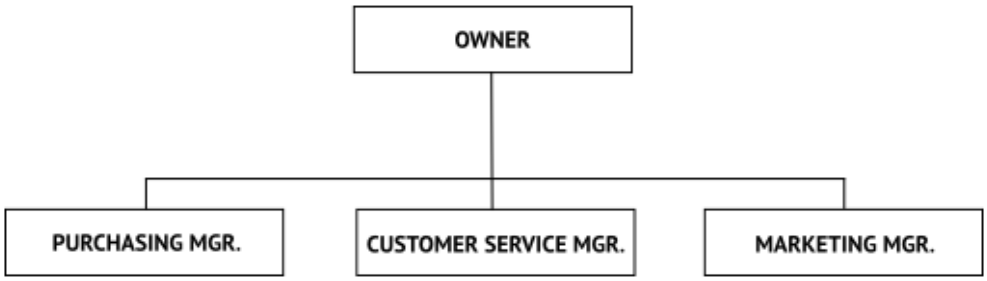
Gambar II.1 Logo Jeger Jersey Indonesia

Makna dari logo pada Gambar II.1 adalah sebagai berikut:

1. Logo berbentuk bulat seperti bola memiliki filosofi bahwa Jeger Jersey Indonesia tidak dapat dipisahkan dengan sepakbola, terutama sepakbola Indonesia.
2. Tulisan "JEGER JERSEY" menggunakan ukuran huruf yang lebih besar dari tulisan lain bertujuan agar *brand* dapat dengan mudah diingat oleh pasar.
3. Tulisan "MMXI" merupakan angka romawi yang berarti 2011, menandakan tahun berdirinya Jeger Jersey Indonesia.

II.1.3. Struktur Organisasi Perusahaan

Gambar II.2 merupakan struktur organisasi dari Jeger Jersey Indonesia.



Gambar II.2 Struktur organisasi Jeger Jersey Indonesia

Sedangkan deskripsi kerja dari setiap bagian pada struktur organisasi di atas adalah sebagai berikut.

a. *Owner*, merupakan pemilik Jeger Jersey Indonesia yang bertugas memberi arahan kepada setiap divisi untuk mencapai tujuan bisnis yang telah ditetapkan. Selain itu, *owner* juga bertugas untuk menentukan berbagai kebijakan seperti penentuan produk dan penetapan harga jual.
b. *Purchasing Manager*, bertugas untuk melakukan pembelian produk dan bekerja sama dengan pemasok barang baku. Selain itu, *purchasing manager* bersinergi dengan *customer service manager* untuk mengelola persediaan produk dan menyediakan produk yang dipesan oleh pelanggan.

c. *Customer Service Manager,* memiliki tugas utama untuk melayani pelanggan termasuk menerima pesanan dan mencatat data transaksi penjualan.
d. *Marketing Manager,* bertanggung jawab untuk memasarkan produk kepada pelanggan. Selain itu, manajer *marketing* bertugas membuat desain promosi dan melakukan analisis pasar.

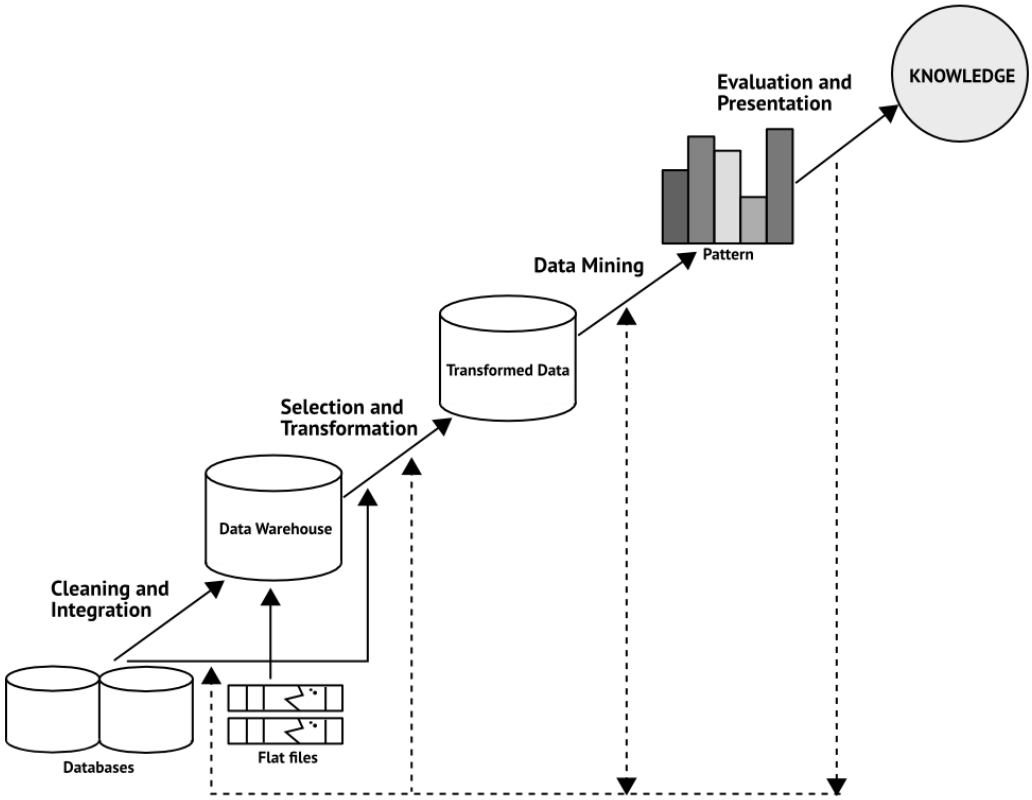
II.2. Landasan Teori

Pada tahap ini dijelaskan beberapa teori yang berkaitan dengan penelitian yang dilakukan. Hal tersebut berguna sebagai dasar pemahaman bagi peneliti dalam menyelesaikan masalah penelitian.

II.2.1. Data Mining

Data mining didefinisikan sebagai proses penemuan pola atau pengetahuan dalam sekumpulan data [5]. Nama lain yang sering digunakan dan memiliki makna yang sama dengan *data mining* adalah *knowledge mining from data*, *knowledge extraction*, *pattern analysis*, *data archaeology*, *data dredging,* dan lain-lain. *Data mining* adalah topik yang melibatkan pembelajaran secara praktis, bukan secara teoritis seperti halnya *machine learning*. *Data mining* berusaha memecahkan masalah dengan menganalisis sekumpulan data yang sudah ada dalam basis data. Sedangkan hasil dari proses dari *data mining* adalah penemuan pengetahuan, informasi, pola, atau model.

Istilah lain dari *data mining* yang banyak digunakan adalah *Knowledge Discovery from Data* (KDD). Selain itu, sebagian lain melihat *data mining* hanya sebagai satu langkah penting pada proses KDD. Gambar II.3 merupakan langkah-langkah dari proses KDD [6].



Gambar II.3 Prosedur *knowledge discovery from data* [6]

Prosedur KDD berdasarkan pada Gambar II.3 adalah sebagai berikut.

a. *Data cleaning*, merupakan proses untuk menghapus data yang mengandung *noise*, data kosong, dan data yang tidak konsisten.
b. *Data integration,* menggabungkan data dari berbagai sumber data.
c. *Data selection,* memilih data yang sesuai dengan analisis yang akan dilakukan.
d. *Data transformation,* mengubah data agar sesuai untuk proses *data mining.*
e. *Data mining,* merupakan proses penerapan metode untuk mengekstrak pola atau pengetahuan dari data.
f. *Pattern evaluation,* mengidentifikasi pola-pola menarik yang dihasilkan.
g. *Knowledge presentation,* menyajikan pengetahuan yang dihasilkan kepada pengguna.

II.2.2. Model Proses CRISP-DM

CRoss-Industry Standard Process for Data Mining (CRISP-DM) adalah metode populer yang digunakan untuk meningkatkan keberhasilan suatu proyek *data mining* [3]. Model proses CRISP-DM terdiri dari enam tahap yang fleksibel. Model ini telah menjadi proses standar proyek *data mining* di banyak perusahaan untuk mendukung keputusan bisnis. Berikut adalah tahap-tahap pada model proses CRISP-DM.

a. Tahap Pemahaman Bisnis

Tahap pemahaman bisnis fokus pada mengidentifikasi tujuan proyek dan kebutuhan dari sudut pandang bisnis. Kemudian mendefinisikan masalah *data mining* dan menyusun rencana untuk mencapai tujuan proyek. Beberapa kegiatan pada tahap pemahaman bisnis adalah sebagai berikut.

1. Melakukan identifikasi tujuan bisnis,
2. Melakukan penilaian situasi,
3. Menentukan tujuan *data mining*, dan
4. Menyusun rencana proyek yang akan dilakukan.

b. Tahap Pemahaman Data

Tahap pemahaman data dimulai dengan pengumpulan data awal, mempelajari data-data tersebut, dan menilai kualitas data. Kemudian dilakukan pendeteksian bagian dari data yang mungkin mengandung informasi tersembunyi. Beberapa kegiatan pada tahap pemahaman data adalah sebagai berikut.

1. Mengumpulkan data awal,
2. Mendeskripsikan data yang diperoleh,
3. Mengeksplorasi data, dan
4. Melakukan verifikasi terhadap kualitas data.

c. Tahap Persiapan Data

Tahap persiapan data digunakan untuk membangun data final dari sekumpulan data mentah. Pada tahap ini dilakukan *data preprocessing* agar data

dapat digunakan untuk proses pemodelan. Kegiatan pada tahap ini mungkin dilakukan beberapa kali dan tidak dalam urutan yang ditentukan. Beberapa kegiatan tersebut adalah sebagai berikut.

1. Melakukan pemilihan data,
2. Melakukan pembersihan data,
3. Melakukan pembangunan data,
4. Mengintegrasikan data dari berbagai sumber, dan
5. Mentransformasikan data sehingga siap untuk diproses.

d. Tahap Pemodelan

Tahap pemodelan merupakan tahap utama pada proses *data mining*. Pada tahap ini diterapkan teknik pemodelan pada data yang telah disiapkan. Beberapa kegiatan pada tahap pemodelan adalah sebagai berikut.

1. Memilih teknik pemodelan yang sesuai dengan data,
2. Menjelaskan prosedur teknik pemodelan yang digunakan,
3. Melakukan penerapan teknik pemodelan, dan
4. Menilai model yang dihasilkan.

e. Tahap Evaluasi

Setelah model terbentuk, perlu dilakukan evaluasi terhadap langkah-langkah yang dilakukan sebelumnya. Hal tersebut dilakukan untuk memastikan model sesuai dengan tujuan bisnis yang ditetapkan. Beberapa kegiatan pada tahap evaluasi adalah sebagai berikut.

1. Mengevaluasi model yang dihasilkan,
2. Mengkaji ulang proses-proses yang dilakukan, dan
3. Menentukan keputusan penggunaan hasil *data mining*.

f. Tahap Penerapan

Tahap ini merupakan tahap implementasi hasil proses *data mining*. Model yang telah dihasilkan perlu diorganisir dan disajikan kepada pengguna. Hal tersebut biasanya berguna untuk proses pengambilan keputusan perusahaan.

Beberapa kegiatan pada tahap penerapan adalah sebagai berikut,

1. Menentukan rencana penerapan hasil *data mining,*
2. Menentukan rencana pengawasan dan pemeliharaan,
3. Membuat laporan akhir, dan
4. Melakukan ulasan terhadap proyek yang telah dilakukan.

II.2.3. Data Prepocessing

Proses pemodelan pada *data mining* memerlukan data yang berkualitas. Namun data pada sebuah perusahaan sangat mungkin untuk tidak akurat, tidak lengkap dan tidak konsisten. Hal tersebut dapat disebabkan oleh banyak faktor, salah satunya adalah kesalahan manusia atau program saat proses memasukkan data. *Data preprocessing* dapat dilakukan untuk memperbaiki kualitas data, sehingga dapat meningkatkan akurasi dan efisiensi hasil *data mining* [6]. Beberapa kegiatan pada *data preprocessing* adalah sebagai berikut.

a. Menangani nilai kosong

Keberadaan nilai kosong pada data adalah masalah yang sering terjadi. Nilai kosong ini akan mempengaruhi hasil analisis data. Nilai kosong pada data biasanya disebabkan oleh kesalahan input atau suatu atribut yang memang tidak memiliki sebuah nilai. Oleh karena itu, nilai kosong pada data perlu ditangani dengan metode yang sesuai. Salah satu cara yang dapat digunakan adalah menghapus data yang memiliki nilai kosong [6]. Setiap objek yang mengandung nilai kosong akan dihapus agar tidak mempengaruhi informasi yang terdapat dalam data. Tabel II.1 merupakan contoh data yang memiliki nilai kosong.

Tabel II.1 Data yang memiliki nilai kosong

Pelanggan	Item	Custom_name	Jumlah_item
A	Jersey Persib, Jersey Sriwijaya FC (Gumbs #17)	1	2
B	Jersey PBR (DEJAN No.punggung 12)	1	1
C	-	0	3
D	Persipura Home & Away Size M, Persib Home Size M 3 buah, Mitra Kukar Away Size S	0	6
E			1

Sedangkan Tabel II.2 merupakan contoh data setelah nilai kosong dihapus.

Tabel II.2 Data setelah nilai kosong dihapus

Pelanggan	Item	Custom_name	Jumlah_item
A	Jersey Persib, Jersey Sriwijaya FC (Gumbs #17)	1	2
B	Jersey PBR (DEJAN No.punggung 12)	1	1
D	Persipura Home & Away Size M, Persib Home Size M 3 buah, Mitra Kukar Away Size S	0	6

b. Menghapus *noise*

Data yang salah, data yang tidak memiliki arti dan *outliers* dapat direpresentasikan sebagai *noise*. Seperti halnya nilai kosong, *noise* juga akan mempengaruhi hasil pemodelan *data mining*. Untuk itu *noise* perlu dihapus agar menghasilkan model yang berkualitas. Cara yang dapat digunakan untuk menghapus *noise* adalah dengan melakukan *smoothing*. Salah satu teknik untuk melakukan *smoothing* adalah dengan menggunakan metode *binning* [6].

Metode *binning* digunakan untuk membagi sekumpulan nilai numerik ke dalam beberapa partisi atau *bin*. *Binning* dimulai dengan mengurutkan setiap nilai pada sebuah atribut. Kemudian setiap nilai dipartisi ke dalam *bin* yang kurang lebih memiliki frekuensi yang sama (*equal-frequency partitioning*). Setelah itu, data pada setiap *bin* diganti dengan nilai batas *bin* terdekat (*smoothing by bin boundaries*). Nilai batas *bin* merupakan nilai minimum dan maksimum pada setiap *bin*.

Sebagai contoh, atribut *jumlah_produk* memiliki 12 nilai yang telah diurutkan yaitu 1, 4, 6, 9, 12, 14, 17, 20, 22, 22, 23, 29. Sedangkan jumlah *bin* yang ditentukan adalah 3. Tabel II.3 merupakan hasil *binning* pada atribut *jumlah_produk*.

Tabel II.3 Hasil *binning* pada sebuah atribut

Bin ke-	Nilai
1	1, 4, 6, 9
2	12, 14, 17, 20
3	22, 22, 23, 29

Untuk melakukan *smoothing* pada atribut *jumlah_produk,* maka semua nilai diganti dengan nilai batas terdekat setiap *bin.* Tabel II.4 merupakan hasil *smoothing* pada atribut *jumlah_produk.*

Tabel II.4 Hasil *smoothing* pada sebuah atribut

Bin ke-	Nilai
1	1, 1, 9, 9
2	12, 12, 20, 20
3	22, 22, 22, 29

c. Pembangunan atribut

Pembangunan atribut (*attribute construction*) merupakan proses pembentukan atribut-atribut baru dari atribut yang sudah ada. Teknik tersebut digunakan untuk membantu dalam proses pembentukan model. Penurunan atribut juga berguna untuk menyeragamkan tipe data atribut, sehingga akan lebih mudah dalam proses pemodelan. Tabel II.5 berisi atribut *jumlah_produk* yang akan diturunkan menjadi beberapa atribut.

Tabel II.5 Data sebelum penurunan atribut

Pelanggan	Jumlah_produk
A	3
B	2
C	1
D	3
E	1

Atribut *jumlah_produk* pada Tabel II.5 akan diturunkan menjadi beberapa atribut dengan tipe data biner. Untuk setiap nilai *v* buat sebuah variabel biner *jumlah_produk_v,* isi dengan nilai 1 jika *jumlah_produk* = *v* dan nilai 0 jika sebaliknya. Tabel II.6 merupakan data hasil penurunan atribut *jumlah_produk* menjadi beberapa atribut biner.

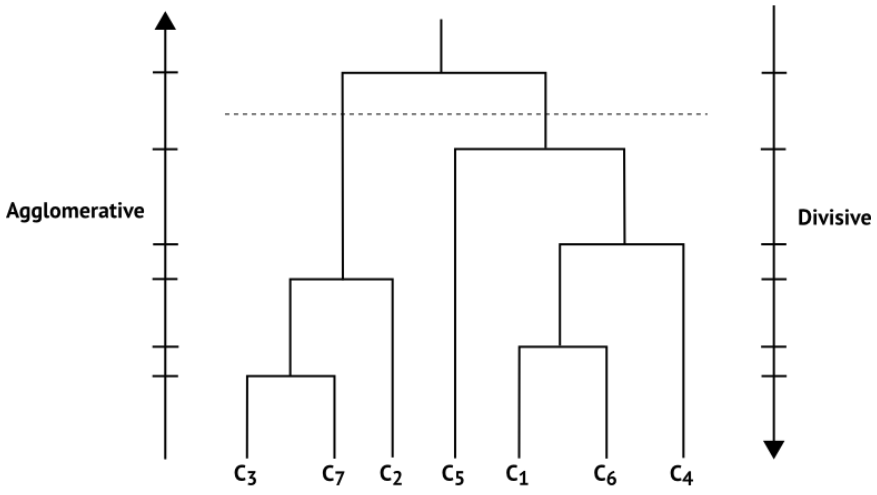
Tabel II.6 Data hasil penurunan atribut

Pelanggan	Jumlah_produk_1	Jumlah_produk_2	Jumlah_produk_3
A	0	0	1
B	0	1	0
C	1	0	0
D	0	0	1
E	1	0	0

II.2.4. Hierarchical Clustering

Hierarchical clustering merupakan metode *clustering* yang dapat melakukan pengelompokkan objek pada data ke dalam sebuah hierarki [7]. Terdapat dua teknik pengelompokkan objek pada *hierarchical clustering* yaitu secara *agglomerative* (*bottom-up*) dan *divisive* (*top-down*). *Hierarchical agglomerative clustering* menggabungkan setiap objek hingga menjadi satu kelompok, sedangkan *hierarchical divisive clustering* memisahkan semua objek pada sebuah kelompok besar menjadi kelompok yang hanya memiliki satu objek.

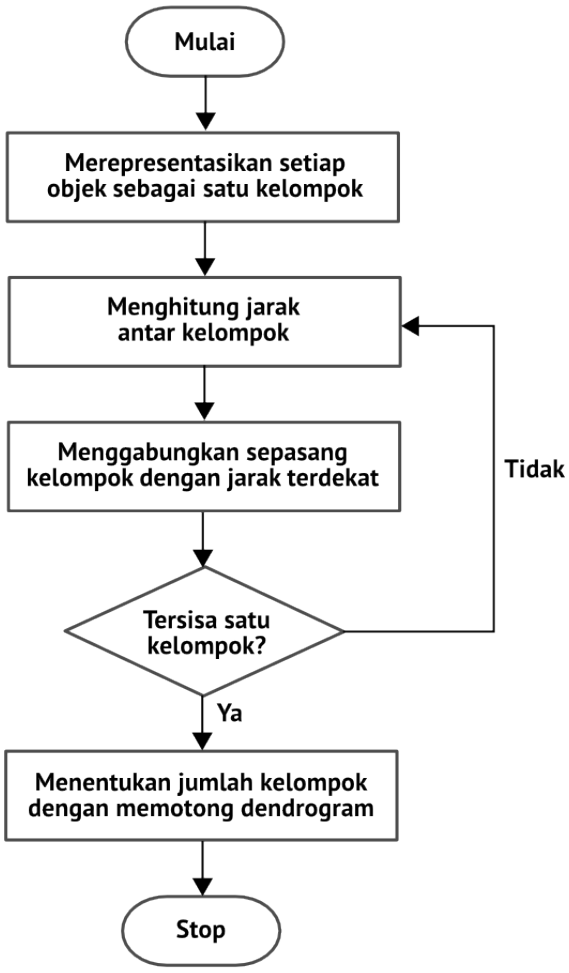
Hasil pengelompokan *hierarchical clustering* dapat direpresentasikan pada sebuah *dendrogram*. *Dendrogram* tersebut merupakan visualisasi struktur pengelompokan data dan dapat memberikan deskripsi yang informatif. Jumlah kelompok yang diinginkan dapat diperoleh dengan memotong *dendrogram* pada suatu jarak tertentu. Terdapat banyak metode dalam menentukan jumlah kelompok yang diinginkan dari sebuah *dendrogram* [8]. Gambar II.4 Merupakan contoh *dendrogram* hasil pengelompokkan menggunakan metode *hierarchical clustering*.



Gambar II.4 *Dendrogram* hasil pengelompokan *hierarchical clustering*

II.2.5. Hierarchical Agglomerative Clustering

Pengelompokkan data menggunakan *hierarchical agglomerative clustering* diawali dengan merepresentasikan setiap objek pada data sebagai satu kelompok, kemudian dilakukan perhitungan jarak (*distance measure*) antar kelompok tersebut. Setelah itu dua kelompok yang memiliki jarak terdekat digabungkan menjadi sebuah kelompok baru. Kemudian jarak antara kelompok yang baru dengan kelompok lain dihitung menggunakan salah satu metode perhitungan jarak antar kelompok (*single linkage*, *complete linkage*, *average linkage*, dll.). Proses perhitungan jarak dan penggabungan dua kelompok dilakukan secara iteratif hingga tersisa satu buah kelompok yang berisi seluruh objek. Gambar II.5 merupakan langkah-langkah dari metode *hierarchical agglomerative clustering.*



Gambar II.5 Langkah-langkah *hierarchical agglomerative clustering*

Perhitungan jarak antar kelompok menjadi faktor penting dalam metode ini. Proses tersebut dilakukan untuk meminimalkan jarak antar objek dalam satu kelompok (*intra-cluster distance*) dan memaksimalkan jarak antara objek dalam satu kelompok dengan objek dalam kelompok lain (*inter-cluster distance*) [8]. Dengan kata lain, perhitungan jarak antara dua kelompok perlu dilakukan untuk mengetahui kemiripan atau kedekatan antar kelompok tersebut.

II.2.6. Jaccard Distance

Jaccard distance merupakan metode yang digunakan untuk menghitung jarak antar setiap kelompok [7]. Metode tersebut digunakan pada sekumpulan data yang memiliki tipe data biner asimetris [9]. Sebelum melakukan perhitungan jarak menggunakan metode *jaccard distance*, perlu dibentuk tabel kemungkinan nilai antara dua kelompok seperti pada Tabel II.7.

Tabel II.7 Tabel kemungkinan nilai antara dua kelompok

x_i \ x_j	**1**	**0**
1	n_{11}	n_{10}
0	n_{01}	n_{00}

Keterangan:

x_i dan x_j : Dua kelompok yang dibandingkan.

n_{11} : Jumlah atribut yang bernilai 1 pada kedua kelompok (*i* dan *j*).

n_{10} : Jumlah atribut yang bernilai 1 untuk kelompok *i* dan bernilai 0 untuk kelompok *j*.

n_{01} : Jumlah atribut yang bernilai 0 untuk kelompok *i* dan bernilai 1 untuk kelompok *j*.

n_{00} *:* Jumlah atribut yang bernilai 0 untuk kedua kelompok (*i* dan *j*).

Sedangkan persamaan untuk menghitung perbedaan jarak antar kelompok menggunakan *jaccard distance* dapat dilihat pada persamaan II.1.

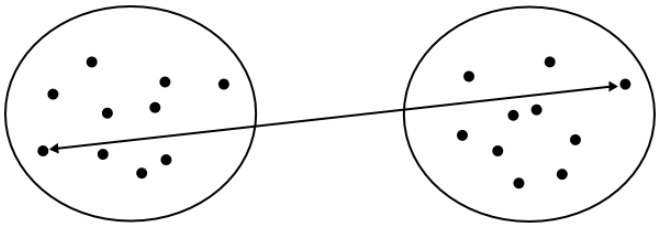
$$dist(x_i, x_j) = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01}} \quad \text{(II.1)}$$

Keterangan:

$dist(x_i, x_j)$: Perbedaan jarak antara dua kelompok (*i* dan *j*), hasilnya akan berkisar diantara nilai 0 sampai 1. Nilai 0 menggambarkan dua kelompok yang identik, dan nilai 1 menggambarkan kelompok yang tidak memiliki kemiripan.

II.2.7. Complete Linkage

Salah satu metode yang dapat digunakan untuk menentukan jarak antar kelompok pada *hierarchical agglomerative clustering* adalah metode *complete linkage.* Pada metode *complete linkage*, jarak antara dua kelompok ditentukan oleh jarak terbesar antara dua objek dalam kelompok yang berbeda [10]. *Complete linkage* membandingkan objek antar kelompok yang paling berbeda di setiap iterasi. Setelah perhitungan jarak dilakukan menggunakan *complete linkage,* dua kelompok yang memiliki jarak terkecil kemudian digabungkan. Gambar II.6 merupakan ilustrasi dari penentuan jarak menggunakan metode *complete linkage.*



Gambar II.6 Ilustrasi penentuan jarak dengan *complete linkage*

Persamaan II.2 merupakan persamaan untuk menentukan jarak antara dua kelompok dengan metode *complete linkage,* dimana $dist(x_i, x_j)$ adalah jarak antara dua objek x_i dan x_j. Sedangkan C_i dan C_j merupakan dua kelompok yang dibandingkan.

$$dist(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} \{dist(x_i, x_j)\} \quad \text{(II.2)}$$

II.2.8. Unified Modeling Language

Unified Modeling Language (UML) adalah notasi berbasis grafik yang digunakan untuk merancang perangkat lunak yang dibangun menggunakan pendekatan berorientasi objek [11]. UML dapat melakukan pemodelan struktur aplikasi, perilaku, arsitektur, proses bisnis dan struktur data. UML merupakan standar terbuka yang dikontrol oleh sebuah perusahaan konsorsium terbuka yaitu *Object Management Group* (OMG). OMG dibentuk untuk membangun standar yang mendukung *interoperability* sistem berorientasi objek. Hingga saat ini versi UML sudah sampai pada versi 2.x. Tabel II.8 merupakan beberapa diagram UML yang akan digunakan pada penelitian.

Tabel II.8 Daftar diagram UML 2.0 yang digunakan pada penelitian

No	Nama Diagram	Fungsi
1	*Use case diagram*	Menggambarkan kebutuhan fungsional sistem dan interaksi antara pengguna dan sistem.
2	*Activity diagram*	Menggambarkan logika prosedural, proses bisnis, dan alur kerja. Selain itu, diagram ini dapat menggambarkan kegiatan yang dilakukan secara bersamaan (paralel).
3	*Sequence diagram*	Menjelaskan perilaku dari sebuah skenario tunggal.
4	*Class diagram*	Menggambarkan tipe dan perilaku objek dalam sistem serta hubungan antar setiap objek.

II.2.9. Silhouette Coefficient

Setelah menerapkan suatu metode untuk menghasilkan segmentasi pelanggan, langkah selanjutnya adalah mengevaluasi segmentasi yang dihasilkan. Evaluasi tersebut dilakukan untuk mengukur kualitas dari hasil pemodelan. Secara umum terdapat dua metode yang dapat digunakan untuk mengukur kualitas dari hasil pemodelan, yaitu metode intrinsik dan metode ekstrinsik [6]. Metode intrinsik digunakan untuk mengukur kecocokan antara segmen yang dihasilkan dengan data yang diolah. Sedangkan metode ekstrinsik digunakan untuk membandingkan antara segmen yang dihasilkan dengan segmen ideal yang ditentukan oleh seorang ahli (*ground truth*).

Jika *ground truth* belum diketahui, metode intrinsik dapat digunakan untuk mengukur kualitas segmen yang terbentuk. Metode intrinsik mengukur seberapa baik pemisahan antara segmen-segmen yang terbentuk dan seberapa *compact* segmen-segmen tersebut. Salah satu teknik yang dapat digunakan pada metode intrinsik adalah dengan menghitung *silhouette coefficient* [6]. Persamaan dari *silhouette coefficient* dari suatu objek **o** adalah

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}, \tag{II.3}$$

dimana

$$a(\boldsymbol{o}) = \frac{\sum_{\acute{\boldsymbol{o}} \in C_i, \boldsymbol{o} \neq \acute{\boldsymbol{o}}} dist(\boldsymbol{o}, \acute{\boldsymbol{o}})}{|C_i| - 1}, \tag{II.4}$$

dan

$$b(\boldsymbol{o}) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{\acute{\boldsymbol{o}} \in C_j} dist(\boldsymbol{o}, \acute{\boldsymbol{o}})}{|C_j|} \right\}. \tag{II.5}$$

Keterangan:

$s(\boldsymbol{o})$: Nilai *silhouette coefficient* dari suatu objek **o**.

$a(\boldsymbol{o})$: Rata-rata jarak antara **o** dan objek-objek lain dalam kelompok yang sama dengan **o**.

$b(\boldsymbol{o})$: Rata-rata jarak minimum dari **o** terhadap semua kelompok yang tidak terdapat **o**.

k : Jumlah kelompok yang terbentuk $(C_i, \ldots, C_k)$.

Nilai dari $s(\boldsymbol{o})$ yaitu antara -1 sampai 1. Jika nilai *silhouette coefficient* mendekati 1, segmen yang mengandung objek **o** adalah *compact* dan objek **o** jauh dari segmen lain. Sebaliknya jika nilai *silhouette coefficient* mendekati -1, objek **o** lebih dekat dengan objek pada segmen lain. Diharapkan setiap objek memiliki nilai *silhouette coefficient* yang mendekati 1, sehingga akan menghasilkan segmentasi yang berkualitas tinggi.