BAB 2
TINJAUAN PUSTAKA

2.1. Data Mining

Data mining yakni proses yang digunakan dalam mengekstrak atau menambang data dalam jumlah yang besar, dengan tujuan untuk mencapai keputusan atau menghasilkan pengetahuan baru. Data mining juga melibatkan kombinasi informasi dari *database* besar, yang mencakup pola, statistik, basis data dan visualisasi informasi.

Data mining memiliki banyak istilah yang dipergunakan dalam memperlihatkan proses data mining, memiliki peran penting, terutama di berbagai organisasi, baik di dunia bisnis ataupun pemerintahan, yang berhadapan dengan banyak informasi beserta pengelolaan basis data. Hal ini sering kali mencakup kebutuhan untuk membangun *Data Warehouse* dalam sekala besar.

Data mining adalah proses untuk mengekstrak data dan mengubahnya dalam bentuk informasi ataupun pengetahuan baru. Data yang diproses sebelumnya sering kali mengandung sifat yang implisit serta dinilai tidak berguna, serta berasal dari volume data yang berada di jumlah yang besar. Ada empat tugas utamanya dalam data mining, yakni mencakupi:

1. *Predictive Modelling*

Predictive Modelling dipergunakan dalam membentuk model yang memprediksi variabel target yang merupakan fungsi akan *explanatory variable*. *Explanatory variable* di sini mencakup keseluruhan dari atribut yang dipergunakan dalam menyusun prediksi, sementara variabel target yakni atribut yang nilainya hendak diprediksi. *Predictive modelling* terbagi atas dua tipe, yakni: *Classification* yang dipergunakan dalam membentuk prediksi terkait nilai variabel target yang bersifat diskrit serta *regression,* yang dipergunakan dalam menghasilkan prediksi dari nilai variabel target yang bersifat *continue* (berkelanjutan).

2. *Association Analysis*

Association Analysis yakni proses guna mengidentifikasi aturan asosiasi yang

memperlihatkan keadaan dari nilai atribut yang tidak jarang muncul dengan bersama-sama pada sekumpulan data.

3. *Cluster Analysis*

Melakukan pengelompokkan objek berdasar pada informasi yang ditemukannya pada data yang menggabarkan objek tersebut beserta hubungannya. Berbagai objek yang mirip akan dikelompokkan di *cluster* yang sama.

4. *Anomaly Detection*

Anomaly Detection yakni metode untuk mendeteksi data yang berbeda dari mayoritas data. *Anomaly* mampu diidentifikasi mempergunakan uji statistik yang menjalankan penerapan akan model distribusi ataupun probabilitas terhadap data.

2.2. *Clustering*

Menurut Widodo (2013:9), *Clustering* ataupun klasifikasi yakni metode yang dipergunakan dalam melakukan pembagian pada sekumpulan data menuju beberapa kelompok yang berdasar pada beragam kesamaan yang telah ditetapkan sebelumnya. *Clustering* kerap dipergunakan menjadi sebuah tahap pertama yang berada di metode data mining. *Cluster* menjadi sekelompok ataupun sekumpulan objek data yang terakait dengan satu sama lainnya pada kelompoknya yang sama serta tak berkaitan dengan objek yang berbeda *cluster*. Jumlah objek yang tergabung pada satu ataupun lebih *cluster* menjadikan objek yang ada di sebuah *cluster* mampu memiliki tingkat kesamaan yang tinggi diantara satu bersama yang lainnya. Objek tersebut terkelompokkan berdasar pada prinsip mengoptimalkan kesamaan objek yang berada di *cluster* yang sama serta mengoptimalkan ketidaksamaan yang berada dalam *cluster* yang tidak sama. Kesamaan objek umumnya didapat melalui berbagai nilai atribut yang menjadi penjelas dalam objek data, yang menjadikan berbagai objek data tersebut umumnya terpresentasikan menjadi suatu titik dalam ruang multidemensi.

Melalui penggunaan clustering, kita mampu mengidentifikasi area dengan kepadatan tinggi, menemukan pola distribusi secara menyeluruh, serta mengungkap hubungan menarik antar atribut data. Dalam data mining, perhatian utama diarahkan menuju pengembangan metode yang mampu menemukan cluster secara

efektif dan efisien pada basis data yang berukuran besar. Beberapa kebutuhan penting dalam clustering untuk data mining mencakup skalabilitas, kemampuan dalam penanganan beragam jenis atribut, kemampuan bekerja dengan data berdimensi tinggi, toleransi terhadap data yang mengandung noise, serta kemudahan interpretasi hasil clustering.

Adapun tujuan dari data *clustering* ini yakni guna meminimalisir fungsi objektif yang diterapkan pada metode *clustering*, yang umumnya berupaya mengurangi variasi antar *cluster*. Secara umum, ada berbagai metode klasifikasi data, dan pemilihan metode *clustering* yang bergantung dengan tipe data serta tujuan dari proses *clustering* itu sendiri.

2.3. Algoritme *K-Means++*

Metode *K-Means++* dipergunakan dalam menangani permasalahan terkait pemilihan pusat *cluster* awal dengan acak yang memicu peningkatan jumlah iterasi. *K-Means++* merupakan penyempurnaan akan algoritma *K-Means* - klasik yang diperkenalkan oleh David Arthur dan Sergei Vassilvitskii pada tahun 2007. Algoritma ini mengandung tujuannya guna memperoleh pemilihan *centroid* awal yang lebih baik daripada pemilihan *centroid* dengan acak pada *K-Means* klasik. *Centroid* awal yang cenderung baik dapat membentuk hasil *clustering* yang juga cenderung optimal serta menekan risiko terjemahnya algoritma pada optimum lokal. *K-Means++* telah banyak diadopsi dan digunakan dalam berbagai aplikasi dan penelitian. Penelitian yang dijalankan oleh Bachem et al. (2016). Memperlihatkan bahwasanya *K-Means++* menawarkan hasil yang cenderung unggul dibanding K-means klasik dalam kasus data dengan distribusi non-convex dan *cluster* dengan ukuran yang sangat berbeda. Untuk meningkatkan kualitas pengelompokan, metode *K-Means++* memastikan pemilihan titik pusat awal yang lebih bijaksana. Berikut adalah langkah-langkah dalam penerapan metode *K-Means++* yakni:

1. Pilih titik *centroid* secara acak

2. Hitung jarak antara setiap titik pada *dataset* dengan *centroid* yang dipilih. Jarak titikx_i pada *centroid* terjauh mampu dilakukan penghitungannya melalui persamaan 2.1.

$$d_i = max_{(j:1\rightarrow m)} || x_i - C_j ||^2 \qquad (2.1)$$

3. Pilih titikx_i yang menjadi *centroid* baru dengan probabilitas maksimum.

4. Ulangi tahapan 2-3 hingga semua titik *centroid* di setiap *cluster*(k) tercapai.

5. Setelah memperoleh *centroid*, langkah berikutnya mengikuti prosedur yang sama seperti pada metode *K-Means*.

2.4. *Euclidean Distance*

Euclidean distance yakni perhitungan jarak antara dua buah titik yang berada di ruang *Euclidean space,* yang pertama kali diperkenalkannya oleh *Euclid*, seorang matematikawan asal Yunani disekitaran tahun 300 SM. *Euclidean Distance* digunakan untuk memperlajari hubungan antara sudut dan jarak. Untuk pengukuran jarak, *Manhattan Distance* dipergunakan dengan Notasi mencakupi berikut:

$$d(x_{i,}\, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.2)$$

Keterangan:

$D\ (x_i, y_i)$ = jarak antara *clustering x* dalam klaster dengan titik *centroid yi*

x_i = bobot ke-i pada klaster yang jaraknya hendak dihitung

y_i = bobot data ke-i di titik *centroid*

n = jumlah data

2.5. Davies Bouldin Index (DBI)

Davies Bouldin Index (DBI) diperkenalkannya oleh David L. Davies dan Donald W. Bouldin (1979) yang dipergunakan dalam menjalankan pengevaluasian pada *cluster*. Pendekatan pengujian nilai DBI berwujud nilai separasi dan kohesi. Kohesi berbentuk jumlah akan kemiripan data pada pusat *cluster* dari *cluster* sementara separasi yakni jarak diantara pusat *cluster* dari *cluster* tersebut.

Berikut ialah tahapan pada evaluasi *cluster* melalui penggunaan metode *Davies Bouldin Index*:

1. *Sum of square within cluster* (SSW) yakni persamaan guna mengetahui matrik kohesi dalam suatu *cluster* ke-i

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{ci} d(X_j, C_j) \quad (2.3)$$

Keterangan :

m_i = jumlah data dalam cluster ke-i

c_i = *centroid* cluster ke-i

$d(x_j, c_i)$ = jarak *euclidean* antara setiap data dengan *centroid*

2. *Sum of square between cluster* (SSB) yakni rumus yang digunakan guna memperoleh pengetahuan terkait nilai pemisah diantara *cluster.*

$$SSB_{i,j} = d(c_i, c_j) \quad (2.4)$$

Keterangan:

$d(c_i, c_j)$ = jarak antara *centroid*

3. Setelah nilai pemisah dan kohesi diperoleh, perhitungan rasio (Ri,j) dilakukan teruntuk memperoleh pengetahuan terkait perbandingan diantara *cluster* ke-i dan *cluster* ke-j, guna menentukan nilai rasio di setiap *cluster*.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_i} \quad (2.5)$$

Keterangan:

SSW_i = *Sum of Square Within Cluster* pada *centroid i*

SSB_i = *Sum of Square Between Cluster* antara data ke-*i* dengan *j* pada cluster yang berbeda

4. *Davies Bouldin Index* (DBI)

Nilai rasio yang diperoleh melalui penghitungan rasio dipergunakan dalam menghitung nilai DBI dengan mempergunakan persamaan berikut:

$$DBI = \frac{1}{k}\sum_{i=i}^{k} max_{i \neq j}(R_{i,j,\ldots,k}) \qquad (2.6)$$

Berdasarkan perhitungan *Davies Bouldin Index* (DBI) tersebut, mampu dihasilkan simpulan bahwasanya makin kecil nilai DBI yang didapat (dengan nilai non negatif≥0)makin baik kualitas *cluster* yang tercipta.

3.1. *Silhouette Coefficient*

Silhouette Coefficient yakni metode evaluasi kualitas *cluster* yang dipergunakan dalam mengamati kualitas beserta kekuatan *cluster*, seberapa baik sebuah objek diletakkan pada sebuah *cluster. Silhouette Coefficient* diperkenalkan oleh Peter J. Rousseeuw dan Michel Kaufman pada tahun 1987. Metode ini termasuk kolaborasi akan metode *cohesion* beserta *separation*. Tahapan perhitungan *Silhouette Coefficient*:

1. Hitung rata-rata jarak pada sebuah data misal *i* bersama seluruh data lain yang berada dalam satu *cluste* yang sama mempergunakan persamaan 2.7.

$$a(i) = \frac{1}{|A|-1}\sum_{J \in A, j \neq i} d(i,j) \qquad (2.7)$$

Keterangan:

$a(i)$ = Perbedaan rata-rata pada data *i* terhadap seluruh data lainnya dalam *cluster A*

i, j = Indeks dari data

$d(i,j)$ = Jarak antara data i dan data *j*

2. Melakukan penghitungan akan rata-rata jarak antara suatu bersama seluruh data di *cluster* lain mempergunakan persamaan 2.8.

$$d\,(i,C) = \frac{1}{|C|}\sum_{J \in C} d(i,j) \qquad (2.8)$$

Keterangan :

$d(i, C)$ = rata-rata jarak pada data (i) ke seluruh data C

C = *Cluster* selain A

3. Pilih nilai jarak minimum mempergunakan persamaan 2.9.

$$b(i) = \frac{min}{c \neq A} d(i, j) \quad (2.9)$$

Keterangan :

$b(i)$ = Nilai jarak rata-rata minimum antara data ke-i dengan seluruh data dalam cluster yang berbeda.

4. Hitung nilai *Silhouette Coefficient* dengan mempergunakan persamaan2.10.

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (2.10)$$

Keterangan :

$a(i)$ = rata-rata jarak antara data i dan semua data lain dalam *cluster A*

$b(i)$ = rata-rata jarak antara data i dan semua data dalam *cluster* berbeda

$s(i)$ = nilai silhouette coefficient

Rentang nilai $s(i)$ berada diantara -1 dan 1, dengan interprestasi nilai sebagai berikut:

$s(i)$ = -1: data ke- i tergolongkan lemah (lebih dekat dengan *cluster* B daripada A)

$s(i) = 0$: data ke- i berada di antara dua klaster (A dan B)

$s(i) = 1$: data ke- i dikelompokkan baik

Berikut adalah *Silhouette Cofficient* menurut Kaufman dan Rousseeuw ditunjukkan dalam tabel 2.1.

Tabel 2.1. Interpretasi Nilai *Silhouette Cofficient*

No	Rentang Nilai SC	Keterangan
1	0,7 < SC≤ 1	*Strong Structure*
2	0,5 < SC ≤ 0,7	*Medium Structure*
3	0,25 < SC ≤ 0,5	*Weak Structure*
4	SC ≤ 0,25	*No Structure*

3.2. Review Artikel

Beberapa referensi dari hasil review artikel hasil penelitian yang dapat dijadikan sebagai dasar penggunaan metode yang dipergunakan pada proposal penelitian ini, mampu diperhatikan melalui tabel 2.2.

Tabel 2.2. Hasil Review Artikel

Landasan Literatur	Metode yang digunakan	Masalah	Hasil Penelitian
Implementasi Data Mining untuk Menentukan Penjualan Alat Perabot Dengan Menggunakan Metode *K-Means Clustering* Pada PT. XYZ.	*K-Means Clustering*	Memiliki keterbatasan dalam menganalisis produk-produk yang dibutuhkan, sehingga menyebabkan kesulitan bagi toko perabot dalam meningkatkan penjualan dan pendapatan perusahaan	Hasil yang di dapatkan bahwasanya terdapat 5 barang yang paling laris serta 5 barang yang kurang laris.

Landasan Literatur	Metode yang digunakan	Masalah	Hasil Penelitian
Implementasi Algoritma *K-Means Clustering* Untuk Pengelompokan Penjualan Produk Pada Online Shop Toko Gizi.	Algoritme *K-Means Clustering*	permasalahan yang mana stok produk yang berputarnya dengan cepat yang memicu pesanan perlu dilakukan penundaannya sebab ketersediaan barang yang sering kosong	Hasil yang didapatkan di *cluster* pertama ataupun Kurang Diminati memiliki 5 barang, *cluster* kedua ataupun Cukup diminati memegang 5 barang serta *cluster* ketiga ataupun sangat diminati memegang 1 barang
Penerapan Data Mining Untuk Koperasi Se-Jawa Barat Menggunakan Metode *Clustering* di Kementerian Koperasi dan UKM.	*K-Means Clustering*	Mengelompokkan data koperasi di Jawa Barat, sehingga dapat dilakukan pengendalian ataupun pemantauan terhadap koperasi yang perlu dipertahankan usahanya serta koperasi yang perlu membentuk peningkatan dalam usahanya	Hasil pengelompokan mempergunakan metode K-Means berdasar akan nilai Modal Sendiri, Modal Luar, beserta Volume Usaha, membentuk tiga klaster, yakni dengan nilai klaster tinggi, sedang, beserta rendah.
Implementasi Data Mining Pada Hasil Penjualan Barang Menggunakan Metode *K-Means*	*K-Means Clustering*	Stok barang yang tidak berkesesuaian	Terdapat pengelompokan produk yang

Landasan Literatur	Metode yang digunakan	Masalah	Hasil Penelitian
Clustering.		pada data yang ada penjualan barang mebel sumber saudara di semarang	menjadi produk sangat laku, laku, dan kurang laku
Metode Algoritma *K-Means* Untuk *Clustering* Data Produk Paling Laku Pada Toko Tono Grosir Plumbon Cirebon.	Algoritme *K-Means Clustering*	toko Tono grosir masih di catat mempergunakan buku ataupun di input dengan manual, terkendala saat melakukan penghitungan akan barang yang sangat laku terjual, laku terjual beserta kurang laku terjual.	Dari hasil perhitungan algoritma k-means berakhir dalam iterasi-4. Ditemukannya 3 *cluster*, yang mana C1 punya 3 data sangat laku, C2 punya 17 data laku, serta C3 punya 12 data kurang laku.
Penerapan Data Mining dengan Metode *K-Means* Untuk Analisis Penjualan di Toko Fashion Hijab Banten.	*K-Means Clustering*	Kurangnya peningkatan strategi pemasaran, data penjualan, serta pengeluaran tidak terduga pada Toko Helai ini tidak disusun secara baik.	Hasil menunjukkan adanya tiga klaster yaitu, sangat laris, laris, dan kurang laris.