BAB II

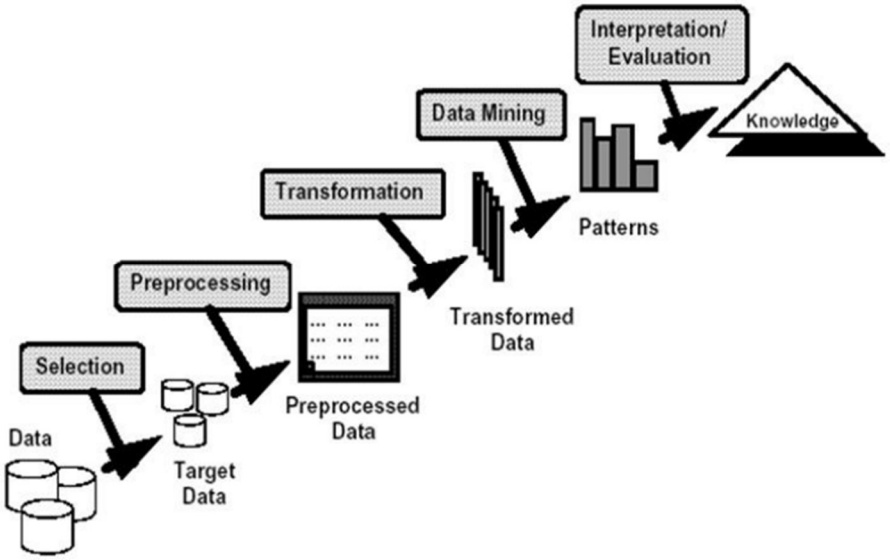
TINJAUAN PUSTAKA

2.1 Data *Mining*

Data *mining* adalah sebuah teknik untuk mengidentifikasi suatu informasi secara menyeluruh terkait pada berbagai *dataset* yang besar untuk mendapat pengetahuan baru (R, Anggraeni, & Enri, 2022). Data *mining* dapat dijelaskan sebagai pengolahan untuk memperoleh informasi dengan menggunakan teknik matematika, statistik, *machine learning* dan *artificial intelligence* yang terikat dengan bermacam data *warehouse* atau *database* berukuran besar yang berguna untuk mengidentifikasi pengetahuan (Muttaqin & Defriani, 2020). Data *mining* bermanfaat untuk mengolah data mentah menjadi kumpulan informasi yang dapat dijadikan sebagai penunjang keputusan secara efektif (Sari, Sudewa, Lestari, & Jaya, 2020).

Data *mining* memiliki beberapa tujuan yaitu *Explonatory*, *Confirmatory* dan *Exploratory*. *Explonatory* bertujuan untuk menjelaskan beberapa kegiatan observasi atau sebuh kondisi. *Confirmatory* bertujuan untuk mengkonfirmasi suatu hipotesis yang telah ada. *Exploratory* bertujuan untuk menganalisis data baru dalam suatu relasi yang janggal (Metisen & Sari, 2015). Berdasarkan fungsinya data *mining* terbagi menjadi enam kelompok yaitu klasifikasi, regresi, *association rule*, *anomaly detection*, *deployment* dan *clustering* (Luchia, Handayani, Hamdi, Erlangga, & Octavia, 2022).

Menurut (Metisen & Sari, 2015) dalam data *mining* terdapat beberapa proses seperti pada Gambar 2.1.



Gambar 2. 1 KDD

Sumber : (Metisen & Sari, 2015)

Berikut adalah penjelasan tahapan data *mining* berdasarkan Gambar 2.1.

1. Data *Selection* yaitu menciptakan himpunan data target dengan memilih himpunan data yang kemudian hasil seleksi disimpan dalam suatu berkas yang terpisah dari basis data operasional.
2. *Pre-processing* atau *Cleaning* yaitu penghapusan *noise* dengan membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data. Dalam tahapan *pre-processing* juga dilakukan normalisasi. Proses normalisasi merupakan proses mengubah skala atau rentang nilai dari data sehingga nilainya dapat dibandingkan atau digunakan secara lebih efektif. Normalisasi data dilakukan pada data numerik untuk menghilangkan perbedaan skala yang mungkin ada antara atribut-atribut tersebut.
3. *Transformation* yaitu proses integrasi pada data yang telah dipilih sehingga data sesuai untuk proses data *mining*.

4. Data *Mining* yaitu pemilihan teknik, metode atau algoritma yang tepat dan sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
5. *Interpretation* atau *Evaluation* merupakan tahap pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.2 Metode *Clustering*

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah *dataset* ke dalam *subset*, sehingga data dalam setiap *subset* memiliki arti yang bermanfaat. Algoritma *clustering* terdiri dari dua bagian yaitu secara *hirarkis* dan secara *partitional*. Algoritma *hirarkis* menemukan *cluster* secara berurutan dimana *cluster* ditetapkan sebelumnya, sedangkan algoritma *partitional* menentukan semua kelompok pada waktu tertentu (Merliana, Ernawati, & Santoso, 2015). *Clustering* dapat diartikan lebih ke arah pengelompokan *record*, pengamatan, atau kasus dalam kelas yang memiliki kemiripan (Suhartini & Yuliani, 2021). Ada beberapa algoritma yang terdapat dalam *clustering* seperti *K-Means* dan *K-Medoids* (Luchia, Handayani, Hamdi, Erlangga, & Octavia, 2022).

2.3 Algoritma *K-Means* dan Metode *Elbow*

Algoritma *K-Means Clustering* merupakan salah satu metode pengelompokkan data nonhierarki (selatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain (Maulida, 2018). Algoritma *K-Means* tergolong sederhana

untuk dapat diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, serta umum penggunaannya dalam praktek (Sembiring, Saifullah, & Winanjaya, 2021).

Algoritma *K-Means* memiliki kelebihan dan kekurangan dalam implementasinya. Kelebihan yang dimiliki yaitu implementasinya yang cukup mudah, serta dapat mengolah data yang cukup besar dengan waktu yang dibutuhkan relatif cepat untuk menjalankannya. Sedangkan, untuk kekurangan algoritma *K-Means* yaitu penentuan nilai *k* secara manual yang berakibat tidak bisa dipastikan nilai *k* mana yang paling terbaik. Selain itu, *K-means* lemah terhadap dimensi yang tinggi (Kasim, Bahri, & Amir, 2021). *Flowchart* dari algoritma *K-means* seperti pada Gambar 2.2.

Gambar 2. 2 *Flowchart* Algoritma *K-Means*
Sumber: (Merliana, Ernawati, & Santoso, 2015)

Langkah-langkah dalam menggunakan algoritma *K-means* menurut Merliana, Ernawati & Santoso sebagai berikut:

1. Tentukan *k* sebagai jumlah *cluster* yang akan dibentuk.
2. Tentukan *k Centroid* (titik pusat *cluster*) awal secara random/acak.

$$v = \frac{\sum_{i=0}^{n} x_i}{n} \quad ; i = 1,2,3, \dots n \tag{1}$$

Dimana;

v = *centroid* pada *cluster*

x_i = objek ke-i

n = banyaknya objek/jumlah objek yang menjadi anggota *cluster*

3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antara objek dengan *centroid* dapat menggunakan *Euclidian Distance*.

$$d(x, y) = |x - y| = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \quad ; i = 1,2,3, \dots n \tag{2}$$

Dimana;

x_i = objek x ke-i

y_i = daya y ke-i

n = banyaknya objek

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling dekat.
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (1).
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama.

Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik (Merliana, Ernawati, & Santoso, 2015). Metode *Elbow* digunakan untuk menentukan jumlah *cluster* dari *dataset* sehingga total dalam variasi *cluster* dapat diminimalkan, penentuan jumlah *cluster* dimulai dari *k=2* dan terus bertambah di setiap langkah dengan ditambah 1 pada nilai *k* (Marisa, Ahmad, Yusof, Fachrudin, & Aziz, 2019). Berikut rumus untuk mendapatkan perbandingannya dengan menghitung *SSE* (*Sum Of Square Error*) dari setiap nilai *cluster*.

$$SSE = \sum_{K=1}^{K} \sum_{X_I \epsilon S_K} \|X_i C_k\|_2^2 \quad (3)$$

Keterangan:

k = jumlah *cluster*

X_i = nilai atribut data ke-i

C_k = nilai atribut titik pusat *cluster* ke-i

Langkah-langkah metode *Elbow* untuk menentukan nilai *k* sebagai berikut:

1. Inisialisasi awal nilai *k*.
2. Menaikkan nilai *k.*
3. Menghitung hasil *sum of square error* dari setiap nilai *k.*
4. Melihat hasil *sum of square error* dari nilai *k* yang turun secara drastis.
5. Menetapkan nilai *k* yang berbentuk siku.

2.4 Penerapan *K-Means* untuk Segmentasi Penduduk Miskin di Indonesia

Pada penelitian yang dilakukan oleh Yuni Radana Sembiring, Saifullah dan Riki Winanjaya yang berjudul "Implementasi Data *Mining* Dalam Mengelompokkan Jumlah Penduduk Miskin Berdasarkan Provinsi Menggunakan Algoritma *K-Means*" menghasilkan 3 *cluster*. *Cluster* pertama berjumlah 9 provinsi, *cluster* kedua berjumlah 22 provinsi dan *cluster* ketiga berjumlah 3 provinsi. Kelebihan pada penelitian ini yaitu penggunaan algoritma *K-means clustering* dapat mengelompokkan data kemiskinan di Indonesia untuk mempermudah melihat data. Kekurangan pada penelitian ini yaitu data yang digunakan dibedakan antara penduduk perkotaan dan pedesaan. Selain itu, nilai *k* sudah ditentukan terlebih dahulu tanpa melakukan perhitungan dan tidak terdapat label pada setiap *cluster* yang dihasilkan (Sembiring, Saifullah, & Winanjaya, 2021).

Penelitian yang dilakukan oleh Nurhafiza Sepriyanti, Rahma Sani Nahampun, Muhammad Hafis Zikri, Isnani Ambarani, dan Akhas Rahmadeyan dengan judul "Penerapan *K-Means Clustering* Untuk Mengelompokkan Tingkat Kemiskinan di Provinsi Riau" menghasilkan 3 *cluster* yaitu *cluster* 1 (kemiskinan tingkat rendah), *cluster* 2 (kemiskinan tingkat sedang), dan *cluster* 3 (kemiskinan tingkat tinggi). Kelebihan pada penelitian ini yaitu penggunaan algoritma *K-means clustering* yang dapat memudahkan mengelompokkan kemiskinan di Provinsi Riau dan mempermudah untuk melihat data. Kekurangan pada penelitian ini yaitu indikator yang digunakan hanya berjumlah 3 indikator dan jumlah *cluster* sudah ditentukan terlebih dahulu tanpa dilakukan perhitungan (Sepriyanti, Nahampun, Zikri, Ambarani, & Rahmadeyan, 2022).

Penelitian yang dilakukan oleh Nabila Nur Fransiska R, Dwi Suci Anggraeni dan Ultach Enri yang berjudul "Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma *K-Means* dengan *Silhouette Coefficient*" menghasilkan 2 *cluster* yaitu *cluster* 0 dengan tingkat kemiskinan tinggi sebanyak 14 wilayah dan *cluster* 1 dengan tingkat kemiskinan rendah sebanyak 13 wilayah. Hasil evaluasi *clustering* menggunakan *Silhoutte Coefficient* menghasilkan nilai sebesar 0.576. Kelebihan pada penelitian ini yaitu dengan penentuan nilai *k* menggunakan metode *Elbow* serta menggunakan metode *Silhouette Coefficient* untuk mengevaluasi *clustering* yang dilakukan. Kekurangan pada penelitian ini yaitu atribut yang digunakan tidak bisa menggambarkan tingkat kemiskinan. (R, Anggraeni, & Enri, 2022).

Dengan studi pendahuluan yang dilakukan dan mengamati beberapa penelitian terkait yang relevan mengenai segmentasi penduduk miskin di Indonesia dengan menerapkan algoritma *K-Means*, penelitian yang diambil memiliki kelebihan yaitu menggunakan data dengan atribut berjumlah 4 yakni Garis Kemiskinan Makanan (GKM), Garis Kemiskinan Non Makanan (GKNM), Indeks Kedalaman Kemiskinan (IkdK) dan Indeks Keparahan Kemiskinan (IkpK). Perhitungan *clustering* juga dilakukan dari data sebelum, saat dan setelah pandemi covid-19. Selain itu, juga dilakukan *clustering* pada data tahun 2021 semester ke-2 dan data 2022 semester ke-1 sehingga penelitian yang dilakukan bersifat dinamis dengan tujuan melihat perubahan clusteringnya. Selanjutnya penentuan nilai *k* dilakukan dengan menggunakan metode *Elbow* serta perhitungan kualitas *clustering* yang dihasilkan menggunakan metode *Silhouette Coefficient*.

2.5 Evaluasi *Clustering*

Dalam mengavaluasi hasil pengelompokan dapat dilakukan dengan validasi *cluster*. Validasi *cluster* berguna untuk mengukur tingkat kebaikan dari *cluster* yang terbentuk, salah satu internal *validation in*dex adalah *Silhouette Coefficient* (Widyadhana, Hastuti, Kharisudin, & Fauzi, 2021). *Silhouette Coefficient* digunakan pada tahap evaluasi ini dengan fungsi untuk menguji kualitas dari *cluster* yang telah dihasilkan. Selanjutnya dapat digunakan mengambil keputusan untuk menentukan digunakan atau tidaknya hasil dari data *mining* tersebut (R, Anggraeni, & Enri, 2022).

Silhouette Coefficient memiliki nilai dengan rentang -1 hingga 1, dengan ketentuan *cluster* yang semakin baik ketika nilai dari rata-rata yang dihasilkan dekat dengan rentang nilai 1 dan buruk apabila dekat dengan rentang -1. Rentang nilai ini dipergunakan untuk menunjukkan korelasi kemiripan dari data yang telah dikelompokkan ke dalam *cluster* (R, Anggraeni, & Enri, 2022). Interpretasi subjektif dari koefisien *silhoutte* yang didefinisikan sebagai lebar *silhoutte* rata-rata maksimal untuk seluruh kumpulan data ditunjukkan pada Tabel 2.1 (Kaufman & Rousseeuw, 2005).

Tabel 2. 1 Pengukuran Koefisien *Silhoutte*

Sumber : (Kaufman& Rousseeuw, 2005)

Koefisien *Silhoutte*	Interpretasi yang diusulkan
0,71-1	Struktur kuat
0,51-0,70	Struktur baik
0,26-0,50	Struktur lemah
≤0,25	Tidak ada struktur substansial yang ditemukan

Menurut (Handoyo, M, & Nasution, 2014) langkah-langkah menghitung nilai *Silhouette Coefficient* sebagai berikut:

1. Hitung rata-rata jarak dari suatu dokumen, misalkan I dengan semua dokumen lain yang berada dalam satu *cluster*.

$$\alpha(i) = \frac{1}{|A| - 1} \sum j \in A, j \neq i \; d(i,j) \tag{4}$$

   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.

$$d(i, C) = \frac{1}{|A|} \sum j \in C \; d(i,j) \tag{5}$$

   Dengan d(i,C) adalah jarak rata-rata dokumen I dengan semua objek pada *cluster* lain C dimana $A \neq C$.

$$b(i) = \min C \neq A \; d(i, C) \tag{6}$$

3. Nilai *Silhouette Coefficient* nya adalah:

$$s(i) = \frac{b(i) - a(i)}{\max (a(i), b(i))} \tag{7}$$

.