BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining merupakan proses menemukan korelasi baru yang bermanfaat, pola dan trend dengan menambang sejumlah repository data dalam jumlah besar, dan menggunakan teknologi pengenalan pola seperti statistic dan teknik matematika (Fatmawati & Windarto, 2018). Data mining disebut juga dengan *Knowledge discovery in database* (KKD) ataupun pengenalan pola. Data mining dapat dibagi menjadi empat kelompok, yaitu model prediksi *(prediction modelling)*, analisis kelompok *(Cluster analysis)*, analisis asosiasi *(association analysis)* dan deteksi anomaly *(anomaly detection)* (Tanjung, Windarto, & Fauzan, 2021).

Data mining telah mendapatkan begitu besar perhatian pada dekade terakhir sehubungan dengan perkembangan hardware yang menyediakan kemampuan komputasi luar biasa yang memungkinkan pengolahan data besar. Asal usul data mining dapat dilihat kembali ke akhir tahun 1980-an pada saat istilah tersebut mulai digunakan, paling tidak dalam kalangan komunitas riset. Data mining didukung oleh kemajuan teknologi, kemampuan CPU dan media yang menyimpan data dalam jumlah besar dan mengolahnya dalam waktu yang lebih cepat (Sudarsono, Leo, santoso, & Hendrawan, 2021).

Tahapan pada proses data mining diawali dari penyeleksian data, proses *cleaning* data, proses transformasi, proses data mining atau proses mencari pola atau informasi dari sebuah data terpilih dan tahap terakhir adalah tahap interpretasi

dan evaluasi yang menghasilkan informasi - informasi baru yang bermanfaat. Secara detail dijelaskan sebagai berikut (Winarta & Kurniawan, 2021) :

1. Data *Selection*

   Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD (Knowledge Discovery in Database). Data hasil seleksi akan digunakan dalam proses data mining.

2. *Pre-processing / Cleaning*

   Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. Transformasi

   Proses transformasi ini adalah proses dimana data yang telah dipilih akan diubah kedalam bentuk dimana data bisa diproses dalam data mining

4. Data Mining

   Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

5. *Interpretation / Evaluation*

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.

2.2 Metode *Clustering*

Clustering merupakan suatu pengelompokkan data mining yang bertujuan untuk melakukan sebuah klasifikasi ataupun memprediksi nilai dari variabel target yang mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi berkelompok-kelompok yang memiliki kemiripan karakteristik sehingga objek yang di dalam *cluster* mirip satu sama dengan yang lainnya dan memiliki perbedaan

dengan objek dari *cluster* yang lain (Tanjung, Windarto, & Fauzan, 2021). Untuk melakukan *clustering* ada beberapa metode yang dapat digunakan, diantaranya metode *K-Means clustering*. Metode ini mampu mengelompokkan data dalam jumlah yang besar dan waktu yang cepat dan efisien. Metode *K-Means* adalah metode klastering berbasis jarak yang membagi data ke dalam sejumlah klaster dan algoritma ini bekerja pada atribut numerik (Sihombing & Sihombing, 2021). Kekurangannya, hasil *clustering* bergantung pada penentuan awal pusat *cluster*, sehingga hasil perhitungan *clustering* dengan metode *K-Means* akan baik jika penentuan pusat *cluster* tepat (Hartanti, 2020).

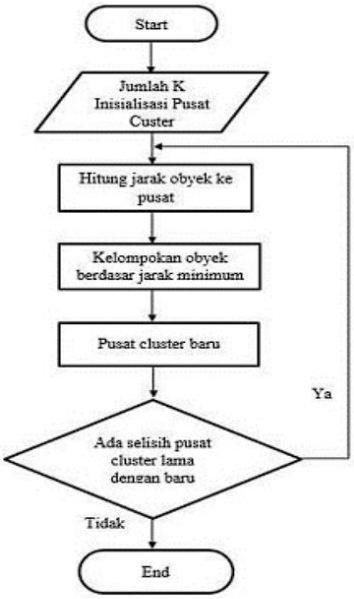
Pada penelitian sebelumnya yang dilakukan oleh Tikaridha Hardianti, 2022 berjudul "Analisis *Clustering* Kasus Covid 19 Di Indonesia Menggunakan Algoritma *K-Means*" menerapkan teknik data mining dengan algoritma *clustering K-Means* dengan mengelompokan kasus Covid 19 di Indonesia yang didapatkan dari website dataset Kaggle. Data yang digunakan sebanyak 16.284 dari tanggal 1 Maret 2020 hingga 9 Juli 2021. Penentuan jumlah *cluster* yang optimal atau validasi *cluster* menggunakan David Boulden index (DBI). *Cluster* yang terbaik ditentukan dari nilai David Boulden Index yang terendah. Hasil penelitian ini diperoleh 3 *cluster* yang terbaik dengan nilai DBI terendah, yaitu sebesar 0,47. *Cluster* 1 terdiri dari 30 provinsi, *Cluster* 2 dan 3 masing-masing 2 provinsi.

2.3 Metode *K-Means* dan *Elbow*

Metode *K-Means* merupakan sebuah metode sederhana untuk membagi suatu kumpulan data dalam suatu angka spesifik dari *cluster*, yaitu k. Metode *K-Means* ditemukan oleh beberapa peneliti dengan disiplin ilmu berbeda-beda yaitu oleh Lloyd (1957, 1982), Forgery (1965), Friedman dan Rubin (1967), dan terakhir

adalah McQueen (1967) (Muningsih & Kiswati, 2018). Metode *K-Means* telah umum diterapkan dalam data mining dan pengenalan pola, dan telah didefinisikan sebagai salah satu pendekatan pengelompokan data mining paling sederhana yang mengimplementasikan fungsi jarak Euclidean (Rahmadhani, 2022). Tujuan proses *clustering* adalah meminimalkan terjadinya *objective function* yang diset dalam proses *clustering*, yang pada umumnya digunakan untuk meminimalisasikan variasi dalam suatu *cluster* dan memaksimalkan variasi antar *cluster* (Muningsih & Kiswati, 2018).

Beberapa keuntungan dari *K-Means* adalah secara singkat diartikan efisien dan cepat. Namun, metode *K-Means* sebagian besar bergantung pada titik data awal dan varian dalam memilih sampel awal yang biasanya diarahkan ke berbagai hasil. Metode *K-Means* selalu menggunakan teknik gradien berdasarkan fungsi tujuan untuk mendapatkan nilai puncak. Fungsi pencarian tren dalam teknik gradien ini terutama diamati dalam proses komputasi di mana seluruh proses akan segera tenggelam ke titik terendah ketika titik fokus *cluster* awal mungkin tidak sesuai yang pada gilirannya menyebabkan pengurangan energi (Rahmadhani, 2022). Berikut Gambar 2.1 merupakan *flowchart* dari algoritma *K-Means*.



Gambar 2. 1 *Flowchart* Algoritma *K-Means*
sumber: Pengelompokan Kabupaten/Kota Berdasarkan Indikator Tingkat Pengangguran Menggunakan Algoritma *K-Means* Clustering (Studi Kasus: Provinsi Jawa Barat), Anggia Arfiani Putrie dan Rangga Sanjaya, 2021 (Putrie & Sanjaya, 2021)

Algoritma dasar *clustering* data menggunakan metode *K-Means* dapat dilakukan dengan cara (Prastyo & Ilfana, 2022):

1. Menentukan jumlah kelompok
2. Menentukan nilai centroid, untuk nilai centroid awal ditentukan secara acak. Selanjutnya untuk menentukan nilai centroid baru pada iterasi digunakan rata-rata dari kelompok ke i untuk variabel ke j dengan rumus sebagai berikut:

$$\overline{V_{IJ}} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj} \qquad (1)$$

Dimana:

$\overline{V_{IJ}}$ = centroid atau rata-rata kelompok ke-I untuk variabel ke j

N_i = jumlah data yang menjadi anggota kelompok ke-i

i, k = indeks dari kelompok

j = indeks dari variabel

X_{kj} = nilai ke-k yang ada didalam kelompok tersebut untuk variabel ke-j

3. Menentukan jarak antara titik centroid dengan setiap titik obyek. Jarak yang digunakan untuk mengukur titik centroid dengan titik obyek adalah jarak Euclid, dengan rumus sebagai berikut:

$$d_{ed} = \sum_{i=1}^{n}(x_i - y_i)^2 \qquad (2)$$

Dimana:

x_i = obyek pengamatan ke i

y_i = centroid ke i

n = banyaknya obyek yang menjadi kelompok

Alokasikan setiap data atau obyek ke *cluster* terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak antar kedua obyek tersebut. Jarang paling dekat antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk ke dalam *cluster* yang mana.

4. Pengelompokan obyek dalam kelompok yang sudah ditentukan berdasarkan jarak yang paling dekat atau minimum.
5. Melakukan iterasi, ulangi langkah 2 dengan menentukan centroid baru dan lanjutkan seterusnya hingga didapatkan hasil tidak terdapat perubahan obyek obyek yang terdapat dalam kelompok.

Salah satu metode yang digunakan untuk menentukan banyak kelompok yang optimal adalah metode *elbow*. 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. Untuk mendapatkan perbandingannya adalah dengan menghitung nilai SSE *(Sum of Square Error)* dari masing-masing *cluster*. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua membentuk suatu siku dalam grafik atau nilainya mengalami penurunan paling besar, maka nilai klaster tersebut adalah yang terbaik (Sihombing & Sihombing, 2021). Berikut rumus SSE pada *K-Means* adalah (Prastyo & Ilfana, 2022):

$$\mathrm{SEE} = \sum_{k=1}^{K} \sum_{x_i \in S_k} (x_i - C_k)^2 \qquad (3)$$

Dimana:

K adalah indeks untuk kelompok dari 1 sampai K, $x_i \in S_k$ adalah obyek ke-1 yang merupakan elemen kelompok S ke-k, C_k adalah centroid pada kelompok ke-k. Adapun langkah-langkah algoritma metode *elbow* dalam menentukan nilai K pada *K-Means* yaitu (Muningsih & Kiswati, 2018):

1. Mulai
2. Inisialisasi awal nilai K
3. Naikkan nilai K
4. Hitung hasil *sum of square error* dari tiap nilai K
5. Melihat hasil *sum of square error* dari nilai K yang turun secara drastis
6. Tetapkan nilai K yang berbentuk siku
7. Selesai

Penelitian sebelumnya Elly Muningsih dan Sri Kiswati, 2018 yang berjudul "Sistem Aplikasi Berbasis Optimasi Metode *Elbow* Untuk Penentuan *clustering* Pelanggan" menggunakan metode *clustering K-Means* dan optimasi metode *Elbow*

dengan mengetahui nilai SEE *(Sum of Square Error)* dihasilkan 3 kelompok pelanggan yang memiliki nilai maksimal atau terbaik.

2.4 Penerapan *K-Means* untuk Pengelompokan Penduduk Pengangguran

Pada penelitian sebelumnya Fadillah Azmi Tanjung, Agus Perdana Windarto, dan M Fauzan, 2021 yang berjudul "Penerapan Metode *K-Means* Pada Pengelompokkan Pengangguran Di Indonesia" yang menggunakan teknik data mining yakni algoritma *K-Means* untuk memecahkan dataset menjadi berkelompok dengan penentuan 2 *cluster* dimana *cluster* 1 adalah kelompok provinsi dengan potensi tertinggi untuk pengangguran dengan hasil 13 provinsi dan *cluster* 2 adalah provinsi dengan hasil pengangguran potensi rendah yaitu 21 provinsi. Pengujian persentase data pengangguran menggunakan *tools Rapidminer* 5.3.

Penelitian sebelumnya Sita Muharni dan Sigit Andriyanto yang berjudul "Penerapan Metode *K-Means Clustering* Pada Data Tingkat Pengangguran Terbuka Tahun 2016-2018 dan 2019-2021" menggunakan metode *clustering* untuk menganalisa perubahan *cluster* TPT 2016-2018 dan 2019-2021 di Indonesia. Dari hasil penelitian didapatkan hanya satu provinsi yang naik *cluster* 1 yaitu Provinsi Riau, pada tahun 2016- 2018 provinsi ini masuk *cluster* provinsi dengan tingkat pengagguran yang tinggi, pada tahun 2019-2021 provinsi Riau naik peringkat menjadi provinsi yang memiliki tingkat pengguran yang rendah. Sebaliknya Provinsi Sumatera Barat turun dari *cluster* 1 sebagai provinsi yang memiliki angka tingkat pengguran rendah menjadi provinsi yang masuk kategori provinsi dengan angka pengangguran yang tinggi pada tahun 2019-2021 (Muharni & Andriyanto, 2022).

Penelitian sebelumnya Akramunnisa dan Fajriani, 2020 yang berjudul "*K-Means Clustering* Analysis Pada Persebaran Tingkat Pengangguran Kabupaten/Kota Di Sulawesi Selatan" indikator yang digunakan adalah Tingkat Pengangguran Terbuka (TPT), Upah Minimum Kabupaten/Kota (UMK) dan laju pertumbuhan Indeks Pembangunan Manusia (IPM). Hasil analisis *K-Means clustering* menunjukkan bahwa 24 kabupaten/kota di Sulawesi Selatan terbagi menjadi dua kelompok, yaitu kelompok tingkat pengangguran tinggi dan rendah. Kelompok dengan tingkat pengangguran tinggi terdiri dari 3 wilayah, yaitu Palopo, Parepare, dan Makassar. 21 wilayah yang lain termasuk pada kelompok dengan tingkat pengangguran rendah. Penelitian ini dapat menambahkan variabel yang mempengaruhi tingkat pengangguran pada wilayah lain sesuai dengan penelitian yang relevan.

Penelitian sebelumnya Mochamad Wahyudi, Lise Pujiastuti, dan Solikhun, 2020 yang berjudul "Penerapan Data Mining Dalam Mengelompokkan Data Pengangguran Terbuka Menurut Provinsi Menggunakan Algoritma *K-Means*" atribut yang digunakan adalah data pengangguran terbuka tahun 2015-2019, menghasilkan *cluster* tinggi dengan jumlah 12 provinsi dan *cluster* rendah dengan jumlah 22 provinsi.

Dengan mengamati beberapa penelitian terdahulu mengenai pengelompokan penduduk pengangguran di Indonesia yang menerapkan algoritma *K-Means* menggunakan data pengangguran yang sudah ada yaitu atribut Tingkat Pengangguran Terbuka (TPT), Upah Minimum Kabupaten/Kota (UMK) dan laju pertumbuhan Indeks Pembangunan Manusia (IPM) sedangkan atribut lainnya yang dianggap penting belum dicantumkan. Oleh karena itu, penelitian ini menggunakan

data dengan atribut berjumlah 4 yaitu jumlah penduduk, Tingkat Partisipasi Angkatan Kerja (TPAK), Indeks Pembangunan Manusia (IPM), dan rata-rata lama sekolah penduduk umur 15 tahun ke atas. Selanjutnya penentuan *k* menggunakan metode *elbow* yang kemudian hasil *cluster*nya akan dievaluasi menggunakan metode *Silhouette Coefficient.*

2.5 Evaluasi *Clustering*

Pada evaluasi *clustering* metode yang digunakan yaitu metode s*ilhouette coefficient.* Metode ini mengevaluasi kualitas sebuah *clustering* dengan menguji seberapa baik *cluster* dipisahkan dan seberapa padat *cluster* tersebut. Biasanya metode ini digunakan ketika tidak terdapat klasterisasi yang ideal sebagai acuan (Orisa & Faisol, 2021). Tahap perhitungan *silhouette coefficient* adalah (Rachman, Goejantoro, & Amijaya, 2020):

1. Menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen yang berada dalam satu *cluster*. Nilai ini disebut a(i) .
2. Kemudian menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen di *cluster* lain. Mengambil nilai terkecil dari semua jarak rata-rata tersebut. Nilai ini disebut b(i) .
3. Kemudian menghitung nilai *silhouette coefficient* dengan menggunakan persamaan:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (1)$$

Dimana:

a(i) = rata-rata dokumen ke-I dengan semua dokumen pada satu *cluster* yang sama

b(i) = rata-rata dokumen ke-I dengan semua dokumen pada *cluster* yang berbeda

S(i) = nilai *Silhouette Coefficient*

Nilai *Silhouette Coefficient* berkisar antara -1 dan 1. Hasil *cluster* dikatakan baik jika nilai *Silhouette Coefficient* adalah 1, berarti dokumen ke-i sudah berada dalam *cluster* yang tepat. Jika nilai *Silhouette Coefficient* adalah 0, maka dokumen ke-i berada di antara dua *cluster*. Jika nilai *silhouette coefficient* adalah -1, artinya struktur *cluster* yang dihasilkan tidak baik, sehingga dokumen ke-i lebih tepat dimasukkan ke dalam *cluster* yang lain. Interpretasi subjektif dari *silhouette coefficient* yang didefinisikan sebagai lebar siluet rata-rata maksimal untuk seluruh kumpulan data ditunjukkan pada Tabel 2.1.

Tabel 2. 1 Pengukuran *Silhouette Coefficient*
(Sumber*: Finding Groups in Data: An Introduction to Cluster Analysis*, Leonard Kaufman dan Peter J. Rousseeuw, 2005 (Kaufman & Rousseeuw, 2005))

Silhouette Coefficient	**Interpretasi yang diusulkan**
0,71-1	Struktur kuat
0,51-0,70	Struktur baik
0,26-0,50	Struktur lemah
≤ 0,250	Tidak ada struktur substansial yang ditemukan

Adapun penelitian sebelumnya Herwinda Kurniadewi, Rijal Abdul Hakim, Mohamad Jajuli, dan Jajam Haerul Jaman, 2022 yang berjudul "Pemetaan UMKM dalam Upaya Pengentasan Kemiskinan dan Penyerapan Tenaga Kerja Menggunakan Algoritma *K-Means*" penelitian ini menghasilkan pengelompokan algoritma *K-Means* menggunakan *Silhouette Coefficient* pada penelitian ini mendapatkan hasil *cluster* terbaik adalah 3 *cluster* yaitu dengan nilai index 0.45 yang mendekati nilai 1.