BAB II
TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining memiliki definisi sederhana sebagai kegiatan dalam melakukan ekstraksi informasi atau pola penting atau menarik dari data dalam *database* besar. *Data mining* adalah penemuan informasi baru dengan mencari pola atau aturan dari sejumlah besar data. Hasil pengolahan data dengan metode *data mining* dapat digunakan untuk mengambil keputusan di kemudian hari. Berdasarkan penjelasan beberapa penelitian sebelumnya, hasil dari proses *data mining* harus berupa informasi baru dan berguna untuk kebutuhan masa depan dan mudah dipahami (Qadrini, 2018). Berikut merupakan tahap-tahapan *data mining* (Sudarsono, et al., 2021)*:*

1. Pembersihan data

   Tahap pembersihan data dilakukan pembersihan agar data-data yang tidak diperlukan lebih baik dihilangkan karena keberadaan data tersebut dapat mengurangi kualitas atau akurasi dari hasil *data mining*.
2. Integrasi data

   Tahap integrasi data dilakukan pada atribut-atribut yang diidentifikasikan sebagai entitas-entitas unik seperti nama, nomor identitas diri dan lain sebagainya. Tahap integrasi data membantu menggabungkan sebuah data berdasarkan entitas-entitas unik.
3. Transformasi data

   Tahap transformasi data dilakukan untuk merubah format data menjadi format khusus sebelum dapat diaplikasikan sehingga didapatkan kualitas dari hasil *data mining*.
4. Aplikasi teknik *data mining*

Tahap aplikasi teknik *data mining* merupakan salah satu bagian dari alur proses *data mining* yang dilakukan untuk membantu melaksanakan *data mining* di bidang tertentu atau untuk data tertentu.

5. Evaluasi pola

   Tahap evaluasi pola dilakukan untuk melihat hasil dari teknik *data mining* dalam menemukan pola-pola yang unik.

6. Presentasi pola

   Tahap presentasi pola dilakukan untuk memformulasikan keputusan dari hasil analisis yang didapat.

2.2 Statistika Deskriptif

Metode mengenai pengumpulan dan penyajian kumpulan data untuk memberikan informasi yang sesuai dengan kebutuhan, yang dikenal sebagai statistik deskriptif. Pada metode ini menggunakan ukuran pemusatan data berupa rata-rata untuk mengetahui karakteristik setiap variabel. Statistika deskriptif bertujuan untuk menemukan karakteristik yang terkait pada setiap variabel (Azkiyah, 2017). Berikut adalah rumus dari rata-rata.

$$\overline{\mathrm{X}} = \frac{\sum_{i=1}^{n} \mathrm{X}}{n} \quad \text{.....(1)}$$

Keterangan :

$\overline{\mathrm{X}}$ = Rata-rata,

X = Jumlah nilai,

n = Banyaknya data.

Standar deviasi merupakan akar dari varians yang digunakan untuk mengevaluasi ukuran keseragaman data setiap variabel. Varians adalah nilai yang menunjukkan tingkat variasi dalam kumpulan data dengan keterangan yang sama dengan rata-rata (Azkiyah, 2017). Berikut rumus standar deviasi adalah sebagai berikut:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X - \bar{X})^2}{n-1}} \quad \text{.....(2)}$$

Keterangan :

$\overline{\mathrm{X}}$ = Rata-rata,

X = Jumlah nilai,

n = Banyaknya data,

S = Standar deviasi.

2.3 Klasterisasi

Klasterisasi (*clustering*) adalah metode yang mengklasifikasikan objek atau data ke dalam kategori yang mirip berdasarkan tingkat kesamaan atau jarak di antara atribut. Tujuan pengelompokan adalah untuk mengelompokkan objek yang mirip ke dalam satu kelompok dan memisahkan objek yang berbeda ke dalam kelompok yang terpisah. Klasterisasi akan mencari pola atau struktur pada data untuk membentuk kelompok dengan karakteristik yang mirip. Atribut ini mungkin atribut numerik atau kategori dari objek yang dianalisis. Klasterisasi pada umumnya menggunakan perhitungan seperti jarak *euclidean* dan lain sebagainya yang digunakan dengan metode pengelompokan untuk mengukur kesamaan antar objek. Klasterisasi berlaku untuk berbagai disiplin ilmu, termasuk *data mining*, analisis, pengenalan pola, dan sistem rekomendasi. Klasterisasi bertujuan utama pengelompokan adalah untuk mengungkap struktur tersirat dalam data, mengidentifikasi pengelompokan yang saling terkait, dan mendapatkan pemahaman dan pendekatan analisis data yang lebih terstruktur dan terorganisir (Tan, et al., 2005).

2.4 Teorema Limit Pusat

Teori teorema limit pusat atau *teorema limit central* adalah salah satu bagian statistika yang paling penting karena mempunyai jangkauan terjauh baik dari segi teori maupun penerapannya dan merupakan sebuah kontribusi modern yang besar, tidak hanya pada statistik tetapi juga pada semua bidang matematika. Teori ini memiliki asumsi-asumsinya langsung bersifat umum, sehingga mudah diterapkan. Teorema limit pusat merupakan teori yang menyatakan bahwa jika ukuran sampel bertambah maka sifat-sifat distribusi *mean* sampel akan semakin mendekati distribusi normal. Hasil pada sampel ini juga mempunyai ciri-ciri yang juga terdapat pada distribusi normal. Semakin banyak data yang diambil maka semakin mendekati distribusi normal (Andini & Trianasari, 2021).

2.5 *One-Way Analyze of Variance* (*One-Way* ANOVA)

One-Way Analysis of Variance (ANOVA) adalah salah jenis dari metode ANOVA yang dimana alat uji ini digunakan untuk menguji perbedaan mean (rata-rata) data lebih dari dua kelompok. Uji *one-way* ANOVA menguji kemampuan dari signifikansi hasil penelitian. Uji ini jika terbukti berbeda dua atau lebih sampel tersebut dianggap dapat mewakili populasi. Uji *one-way* ANOVA dilakukan bertujuan dan pengujian *one-way* ANOVA adalah untuk mengetahui apakah terdapat perbedaan antara berbagai kriteria dengan hasil yang diinginkan. Syarat dalam melakukan uji *one-way* ANOVA pada saat pengambilan sampel yang dilakukan secara *random* terhadap beberapa (> 2) kelompok yang independen, di mana nilai pada satu kelompok tidak tergantung pada nilai di kelompok lain. Uji *one-way* ANOVA harus memenuhi beberapa asumsi sebagai berikut (Palupi & Prasetya, 2022).

1. Sampel terdiri dari kelompok yang independen.
2. Varian antar kelompok harus homogen.
3. Data masing-masing kelompok berdistribusi normal.

2.6 Algoritma *K-Prototype*

K-Prototype adalah algoritma dasar untuk melakukan pengelompokan gabungan data numerik dan kategori. Algoritma ini menggabungkan algoritma *K-Means* dengan algoritma *K-Modes.* Algoritma *K-Prototype* adalah teknik pengelompokan berdasarkan partisi. Algoritma ini berasal dari algoritma *K-Means* untuk mengelola klaster data dengan tipe atribut numerik dan atribut kategori. Metode ini mempertahankan efektivitas *K-Means* dalam menangani kumpulan data besar dan dapat diterapkan pada data numerik dan kategorikal. Inovasi inti algoritma *K-Prototype* adalah ukuran kesamaan antara objek dan *centroid*-nya (*prototype*) (Subhan, et al., 2022). Berikut tahap-tahap dalam menggunakan algoritma *K-Prototype* (Qadrini, 2018).

1. Tahap 1

   Menentukan nilai *k* dengan inisial klaster $z_1, z_{2,\ldots,}z_k$ secara acak (*random*) dari *n* ke sebuah titik $\{ x_1, x_{2,\ldots,}x_n\}$.
2. Tahap 2

Melakukan perhitungan jarak pada seluruh data *point* pada dataset terhadap klaster awal yang telah ditentukan dengan persamaan $d(X_j, Z_i = \pi r^2 = \left(\sum_{I=1}^{m_r}\left({x_{jl}}^r - {z_{il}}^r\right)^2\right)^{\frac{1}{2}}$, selanjutnya melakukan alokasi titik data ke dalam klaster yang memiliki jarak terdekat dengan objek yang diukur (*prototypes*).

3. Tahap 3

   Melakukan perhitungan untuk titik pusat klaster yang baru setelah semua objek dialokasikan, selanjutnya melakukan relokasi pada semua titik data pada dataset terhadap *prototype* yang baru.

4. Tahap 4

   Melakukan pemeriksaan ulang jika titik pusat klaster tidak berubah atau telah konvergen menandakan bahwa proses algoritma telah selesai tetapi jika titik pusat dari klaster masih mengalami perubahan secara signifikan menandakan bahwa proses akan kembali pada tahap 2 dan 3 yang bertujuan untuk mendapatkan iterasi maksimum atau objek tidak mengalami perpindahan.

2.7 *Preprocessing Data Mining*

Pengolahan *data mining* memerlukan sebuah tahap yaitu *preprocessing data mining* untuk menemukan data yang memiliki nilai yang hilang (*missing value*), distorsi nilai, tidak tersimpannya nilai (*misrecording*), *sampling* yang tidak cukup bagus dan sebagainya. Permasalahan tersebut memerlukan sebuah solusi untuk meningkatkan kualitas pengolahan data dengan melakukan sebuah penyiapan data (*preprocessing*). Hal yang menjadi penyebab dari kurang baiknya kualitas dari sebuah data mentah adalah dikarenakan sebuah kesalahan dalam penyimpanan dan pengukuran, tetapi penyebab lain adalah karena tidak adanya nilai yang mewakili nilai yang tersedia (Qadrini, 2018).

2.8 Evaluasi Pengelompokan

Parameter dalam mengetahui kelayakan dari hasil pengelompokan dilakukan evaluasi sebagai berikut.

2.8.1 Uji Validitas Kelompok

Permasalahan utama dalam analisis kelompok adalah menentukan jumlah kelompok oleh peneliti karena belum memiliki dasar yang kuat mengenai jumlah kelompok terbaik. Uji validitas kelompok dilakukan dalam menyelesaikan permasalahan sebelumnya perlu dilakukan uji validitas kelompok dalam menguji hasil dari analisis kelompok secara kuantitatif sehingga dapat dihasilkan kelompok yang optimal (Qadrini, 2018). Kelompok optimal adalah sebuah kelompok yang memiliki jarak yang padat antar individu dalam kelompok dan terisolasi dari kelompok lainnya dengan baik (Jain & Dubes, 1988). Adapun cara dalam menentukan validitas suatu data diukur dengan ukuran internal dan ukuran eksternal (Steinbach, et al., 2000).

2.8.1.1 Ukuran Internal

Ukuran internal digunakan dalam mengukur kelompok data yang terbentuk tanpa pertimbangan informasi dari luar (Steinbach, et al., 2000).

1. Indeks *Global Silhoutte*

   Adapun mendapatkan indeks kelompok *S(i)* adalah berikut

   $$S(i) = \frac{b(i)-a(i)}{\max(a(i).b(i))} \quad \text{.....(4)}$$

   Keterangan :

   a(i) = Rata-rata perbedaan dari *i*-objek dengan semua objek lain pada kelompok yang sama,

   b(i) = Objek pada kelompok lain (di kelompok terdekat).

   Nilai yang paling besar dari indeks *global silhouette* atau rata-rata koefisien *silhouette* menentukan jumlah kelompok terbaik yang akan diambil menjadi kelompok optimal.

   Adapun rumus *global silhoutte* adalah sebagai berikut.

   $$GS_n = \frac{1}{n}\sum_{i=1}^{n} S(i) \quad \text{.....(5)}$$

   Keterangan :

   GS_n = Indeks *global silhouette,*

   S(i) = *Silhoutte* kelompok ke-*i.*

2. Indeks *Dunn*

Adapun indeks *dunn* dirumuskan sebagai berikut.

$$D = \min_{1 \le l \le n} \left\{ \min_{1 \le j \le n, i \ne j} \left\{ \frac{d(c_i, c_j)}{\max_{1 \le k \le n} (d'(c_k))} \right\} \right\} \quad \text{.....(6)}$$

Keterangan :

D = Indeks validasi *dunn*,

$d'(c_k)$ = Jarak dalam kelompok (c_k).

Nilai terbesar dari *D* dipilih menjadi jumlah kelompok optimal (Bolshakova & Azuaje, 2003).

2.8.1.2 Ukuran Eksternal

Ukuran eksternal digunakan dalam mengukur tingkat kecocokan kelompok yang telah terbentuk dengan informasi kelas data. Dalam ukuran validasi eksternal terdapat ukuran kualitas eksternal adalah *entropy* (Steinbach, et al., 2000). Dalam menentukan *entropy* melakukan perhitungan kelas distribusi untuk setiap kelompok. Selanjutnya hitung p_{ik} peluang bahwa kelompok *k* memuat anggota kelas ke-*i*, dengan $p_{ik} = \frac{n_{ik}}{n_k}$. Adapun nilai *entropy* setiap kelompok adalah sebagai berikut (Qadrini, 2018).

$$E_k = -\sum_i p_{ik} \log(p_{ik}) \quad \text{.....(7)}$$

Selanjutnya jumlah total *entropy* dinyatakan dengan persamaan sebagai berikut.

$$E = \sum_{k=1}^{g} \frac{n_k \times E_k}{n} \quad \text{.....(8)}$$

Keterangan :

E = *Entropy*,

Nik = Banyaknya anggota pada kelas ke-*i* yang terdapat pada kelompok *k*,

Nk = Banyaknya anggota kelompok ke-*k*,

g = Banyaknya kelompok.

2.9 *General Linear Model* (GLM)

General Linear Model (GLM) adalah model statistik yang dapat digunakan untuk menganalisis data yang berskala interval atau rasio. GLM merupakan metode

hasil perkembangan dari model linear dengan asumsi prediksi yang memiliki pengaruh linier tetapi tidak melakukan asumsi untuk distribusi tertentu dari variabel *respond* (Caraka, et al., 2018). GLM merupakan metode yang didukung dengan perhitungan *Root Mean Square Error* (RMSE) yang memiliki fungsi untuk mendapatkan pendekatan berdasarkan tingkat kesalahan hasil perkiraan, semakin kecil (mendekati 0) nilai RMSE menunjukkan bahwa semakin akurat nilai perkiraannya (Hamdanah & Fitrianah, 2021).