BAB II
LANDASAN TEORI

2.1 *Data Mining*

Data mining merupakan proses pencarian pola atau informasi menggunakan teknik atau metode tertentu dalam suatu data. Teknik dan metode dalam *data mining* sangat variatif. Pemilihan metode atau algoritma yang tepat sangat berpengaruh pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan (Mardi, 2017). KDD merupakan proses menentukan informasi berguna yang ada di dalam data (Ramadana, Satyahadewi, dan Perdana, 2022.)

Data mining menggunakan beberapa teknik dengan tujuan memperoleh pengetahuan dan informasi yang berkaitan dengan *database* besar. Data yang memiliki ukuran lebih besar biasanya diolah menggunakan *data mining*, kemudian mencari pola atau tren pada data yang sesuai dengan tujuan penerapan *data mining*. Selanjutnya hasil dari pengolahan *data mining* tersebut digunakan untuk pengambilan keputusan maupun hasil prediksi analisis yang dibutuhkan (Mulyati, Yulianti, dan Saifudin, 2017).

Data mining merupakan serangkaian proses yang dibagi menjadi beberapa tahap. Menurut Han, Kamber, dan Pei (2012) terdapat tujuh tahapan dalam *data mining* itu sendiri di antaranya adalah:

1. *Data cleaning*

   Pembersihan data atau *data cleaning* pada *data mining* bertujuan untuk menghilangkan *noise data* atau data yang tidak konsisten. Biasanya data yang diperoleh dari hasil pengujian atau *database* perusahaan berisi data yang tidak lengkap seperti adanya *missing data,* data yang tidak valid maupun kesalahan ketik.
2. Integrasi Data (*Data Integration*)

   Integrasi data adalah penyatuan data dari beberapa sumber.
3. Seleksi Data (*Data Selection*)

   Seleksi data adalah proses pemilihan data yang relevan dari suatu *database*. Data yang terdapat dalam *database* seringkali tidak digunakan seluruhnya.

Oleh karena itu, hanya data yang cocok untuk analisis yang diambil dari *database*.

4. Transformasi Data (*Data Transformation*)

   Fungsi dari transformasi data dalam *data mining* yaitu mengubah data sesuai format untuk dianalisis. Sebelum digunakan, beberapa teknik dalam data *mining* biasanya membutuhkan format data tertentu.

5. Penggalian Data (*Data Mining*)

   Data mining merupakan proses pencarian pola atau informasi dari suatu data menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data *mining* sangat bervariasi.

6. Evaluasi Pola (*Pattern Evaluation*)

   Pola informasi yang diperoleh dari proses *data mining* perlu disajikan ke dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini adalah bagian dari proses KDD yang disebut dengan *interpretation*.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

   Knowledge Presentation adalah proses visualisasi dan representasi pengetahuan dengan tujuan menyajikan pengetahuan dan informasi yang telah digali kepada pengguna.

Menurut Mustafa, Ramadhan, dan Thenata (2018) terdapat enam fungsi *data mining* yaitu:

1. Deskripsi (*description*), merupakan fungsi yang memberikan cara penggambaran pola dan kecenderungan dari sejumlah data berskala besar secara singkat, diantaranya Metode *Neural Network* dan *Decision Tree*
2. Estimasi (*estimation*), merupakan fungsi yang dapat memperkirakan suatu nilai yang belum diketahui kebenarannya, diantaranya Metode *Simple Linear Regression, Point Estimation* dan *Multiple Regression*.
3. Prediksi (*prediction*), merupakan fungsi yang dapat memperkirakan suatu nilai di masa depan, diantaranya Metode *Decision Tree, K-Nearest Neighbor* dan *Neural Network*.
4. Klasifikasi (*classification*), fungsi ini adalah proses penemuan model yang dapat membedakan kelas data dengan tujuan memprediksi suatu kelas dari

objek yang labelnya tidak diketahui, diantaranya adalah Metode *Decision Tree*, *Neural Network*, Algoritma C4.5 dan Naïve Bayes.

5. Pengelompokkan (*clustering*), adalah fungsi yang digunakan untuk mengelompokkan data untuk identifikasi data dengan karakteristik tertentu, diantaranya *Hierarchical Clustering* dan *K-Means*.
6. Asosiasi (*association*) atau disebut juga analisis keranjang pasar yang digunakan untuk mengidentifikasi item-item produk yang mungkin dibeli konsumen bersamaan dengan produk lain, diantaranya Metode *Apriori* dan *FP-Growth*.

2.2 Klasifikasi

Klasifikasi merupakan suatu metode dalam *data mining* yang digunakan untuk menemukan model yang menjelaskan data. Menurut Han, Kamber, dan Pei (2012) klasifikasi merupakan proses menemukan model yang dapat menjelaskan dan membedakan kelas data dimana struktur klasifikasinya digunakan untuk memprediksi label kategorik.

Tujuan model klasifikasi dalam *data mining* adalah sebagai berikut:

a. Pemodelan Deskriptif

   Digunakan sebagai alat yang dapat menjelaskan perbedaan antara objek dengan kelas-kelas yang berbeda.

b. Pemodelan Prediktif

   Digunakan sebagai alat yang dapat memprediksi suatu label kelas yang *record*-nya belum diketahui.

2.3 *Decision Tree*

Decision tree atau pohon keputusan merupakan salah satu metode klasifikasi yang popular karena dapat dengan mudah diinterpretasi (Setio, Saputro, dan Winarno, 2020). *Decision tree* berguna dalam mengeksplorasi data dan menemukan hubungan beberapa variabel input dengan sebuah variabel target. Karena *decision tree* memadukan antara eksplorasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan bahkan

ketika dijadikan sebagai model akhir dari beberapa teknik lain (Kamagi dan Hansun, 2014).

Decision tree merupakan diagram alir seperti pohon di mana setiap simpulnya (*node*) menunjukkan suatu *test* pada suatu atribut, hasilnya diwakili oleh setiap cabang pengujian, dan kelas-kelasnya diwakili oleh simpul daun (*leaf node*). *Decision tree* digunakan untuk mengeksplorasi data yang telah melewati tahap *preprocessing* dan menemukan model yang tersembunyi dari data dengan sebuah target variabel, sehingga dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan *record* yang lebih kecil dengan memperhatikan variabel tujuannya (Jusia, 2018).

2.4 Algoritma C5.0

C5.0 merupakan algoritma dalam *data mining* yang digunakan untuk membangun pohon keputusan. C5.0 merupakan penyempurnaan algoritma sebelumnya yaitu ID3 dan C4.5 yang dibentuk oleh Ross Quinland tahun 1987 (Umma, Warsito, dan Maruddani, 2021). C5.0 memberikan tingkat akurasi terbaik dan waktu eksekusi yang lebih sedikit dibandingkan dengan algoritma klasifikasi lain (Tanti, Sirait, dan Andri, 2018). Rumus yang digunakan dalam algoritma C5.0 adalah sebagai berikut:

a. Sebagai dasar untuk membentuk *node* atau akar dan cabang dari pohon keputusan, *gain ratio* dihitung sebagai berikut:

$$Gain\ Ratio = \frac{Gain(S, A)}{SplitInfo(S, A_i)} \quad (2.1)$$

dengan $Gain(S, A)$ adalah ukuran efektifitas atribut independen A dalam mengklasifikasikan data dan $SplitInfo(S, A_i)$ menyatakan informasi potensial pada atribut independen A kelas ke-i.

b. $SplitInfo(S, A_i)$ dan $Gain(S, A)$ dihitung sebagai berikut:

$$SplitInfo(S, A_i) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (2.2)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropy(S) \quad (2.3)$$

Dengan S adalah jumlah seluruh sampel, A_i adalah atribut independen ke-i, S_i adalah jumlah seluruh sampel untuk kategori ke-i, dan k adalah banyaknya kategori pada atribut independen A.

c. Rumus $Entropy$ adalah sebagai berikut:

$$Entropy(S) = -\sum_{i=1}^{n} P_i log_2 P_i \qquad (2.4)$$

Dengan S adalah jumlah seluruh sampel, n adalah jumlah kelas pada atribut dependen, dan P_i adalah proporsi banyaknya data kelas ke-i pada data.