BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine leaming untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan terkait dari berbagai database besar. Pendekatan dasar dalam data mining adalah merangkum data dan mengambil informasi yang bermakna dan yang sebelumnya tidak diketahui (Suntoro, 2019).

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Amalia, 2018).

2.2 Teknik-Teknik Data Mining

Beberapa teknik dan sifat data mining adalah sebagai berikut :

1. Klasterisasi. Adalah mempartisi data-set menjadi beberapa sub-net atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set property yang di share bersama, dengan tingkat similaritas yang tinggi dalam suatu kelompok yang rendah. Disebut juga dengan "*unsupervised learning*".

2. Regresi. Adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai adari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
3. Klasifikasi. Adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut juga dengan "*supervised learning*".
4. Kaidah Asosiasi (*association rule*). Adalah mendeteksi kumpulan atribut-atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut.

2.3 Tipe Data Mining

1. Tipe Data Numerik adalah tipe data yang diperoleh dengan cara pengukuran dimana jarak dua titik pada skala sudah diketahui.
2. Tipe Data Kategorial adalah tipe data yang diperoleh dengan cara kategorisasi atau klasifikasi.
3. Tipe Data Rentang Waktu adalah tipe yang diperoleh dengan cara menunjukan beberapa objek yang berbeda.

2.4 Tahap-Tahap Data Mining

Menurut Tan (2004), serangkaian proses tahapan memiliki tujuh tahapan yaitu:

1. Pembersihan Data (*data cleaning*)

   Pembersihan data adalah proses untuk menghilangkan data-data yang tidak relevan. Data-data yang dibuang terkadang dibandingkan terlebih dahulu

dengan hipotesa yang telah dibuat. Sehingga pada proses selanjutnya dapat dengan mudah menemukan hasil yang diinginkan.

2. Integrasi data (*data integration*)

   Integrasi data merupakan proses dalam menggabungkan data dari berberapa database kedalam satu database baru. Tidak sedikit data yang dibutuhkan diambil dari berbagai database atau teks file.

3. Seleksi data (*data selection*)

   Data yang sudah ada di database seringkali tidak semuanya dibutuhkan, maka dari itu dibutuhkan penyeleksian data untuk data yang benar-benar dibutuhkan dalam proses selanjutnya.

4. Transformasi data (*data trasnformation*)

   Data digabung atau diubah sesuai dengan proses yang digunakan dalam data mining. Karena beberapa format data mining membutuhkan format data yang khusus dalam pemrosesannya.

5. Proses mining

   Adalah proses menggali data dari sebuah database atau kumpulan data untuk memperoleh informasi yang tersembunyi dari data yang diolah

6. Evaluasi Pola (*pattern evaluation*)

   Dalam proses ini adalah hasil dari teknik data mining berupa pola-pola yang akan diujia pada hipotesa yang sudah dibuat sebelumnya. Sehingga akan

memperoleh kesimpulan-kesimpulan yang mendekati hasil atau hipotesa untuk proses selanjutnya.

7. Presentasi pengetahuan (*knowlegde presentation*)

Ini termasuk dalam langkah akhir dari data mining dalam tahap ini saatnya untuk mempresentasikan hasil yang telah di lakukan dengan mengimplementasikan analisis yang didapat. Sehingga akan memperoleh kesimpulan real.

2.5 K-Means *Clustering*

K-Means merupakan salah satu metode *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda dikelompokan ke dalam cluster yang lain (Priyatman, 2019).

Algoritma K-Means menggunakan proses secara berulang-ulang untuk mendapatkan basis data cluster. Dibutuhkan jumlah cluster awal yang diinginkan sebagai masukan dan menghasilkan titik centroid akhir sebagai output. Metode *clustering* K-Means digunakan untuk mengelompokkan data kedalam kluster dengan melihat nilai centroid yang sudah ditentukan (Meilia, 2019).

Clustering K-Means bekerja dengan cara membagi objek data ke dalam beberapa kelompok/cluster yang berbeda sesuai dengan ukuran kesamaan dari data-data tersebut, sehingga untuk objek data yang berada dalam cluster mempunyai

tingkat kesamaan terbesar sedangkan untuk objek data antar cluster yang berbeda mempunyai tingkat kesamaan terkecil (Xie dkk, 2020).

2.6 CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) suatu standarisasi pemrosesan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur dan terdefinisi dengan jelas dan efisien. Selain menerapkan suatu model dalam proses penambangan data, pemilihan algoritma sangat mempengaruhi terhadap komparasi kinerja metode data mining (Msy dkk, 2021).

Model proses CRISP-DM memberikan gambaran tentang siklus hidup proyek data mining. CRISP-DM memiliki 6 tahapan yaitu :

1. *Business understanding*

   Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun.

2. *Data Understanding*

   Tahapan untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasaan (*summary*) dan mengidentifikasi potensi masalah dalam data.

3. *Data Preparation*

Tahap ini adalah untuk memperbaiki masalah dalam data, kemudian membuat variabel derived. Tahap ini merupakan tahap yang sering ditinjau kembali saat menemukan masalah pada saat pembangunan model. Sehingga dilakukan iterasi sampai menemukan hal yang cocok dengan data. Tahap sampling dapat dilakukan disini dan data secara umum dibagi menjadi dua, data training dan data testing.

4. *Modeling*

Pada tahap ini dilakukan metode statistika dan *Machine Learning* untuk penentuan terhadap teknik data mining, alat bantu data mining, dan algoritma data mining yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik *data mining* tertentu, dapat kembali ke tahap *data preparation*. Beberapa *modeling* yang biasa dilakukan adalah *classification, scoring, ranking, clustering, finding relation,* dan *characterization*.

5. *Evaluation*

Melakukan interpretasi terhadap hasil dari data mining yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

6. *Deployment*

Perencanaan untuk *Deployment* dimulai selama *Business understanding* dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengkonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional.

2.7 Interpretasi Data

Interpretasi data adalah proses meninjau data sampai pada kesimpulan yang relevan dengan menggunakan berbagai metode analisis. Analisis data membantu penelitian dalam mengkategorikan, memanipulasi, dan meringkas data untuk mendapatkan hasil kesimpulan. Langkah-langkah dalam interpretasi data adalah :

1. Pengumpulan Data

   Langkah pertama dalam interpretasi data adalah mengumpulkan semua data yang relevan. Langkah ini bertujuan untuk menganalisis data secara akurat dan tanpa bias.

2. Pengembangan

   Tahap ini adalah ringkasan dari data yang diperoleh, lalu memproses data secara menyeluruh untuk mengidentifikasi tren, pola, atau perilaku.

3. Menarik Kesimpulan

Setelah tahap pengembangan, kemudian menarik kesimpulan berdasarkan tren yang ditemukan untuk mengevaluasi penggunaan data lebih lanjut.

4. Rekomendasi

Tahap terakhir adalah memberikan rekomendasi berdasarkan hasil kesimpulan data yang didapat.