BAB II

LANDASAN TEORI

2.1 Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database [7]. Data mining adalah proses yang menggunakan teknik statik, matematika, kecerdasan buatan dan machine learning untuk mengekstraksian dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [8]. Data Mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut [9].

Selain defenisi di atas beberapa defenisi juga diberikan seperti, data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Menurut Pramudiono data mining adalah menganalisis secara otomatis dari data yang berjumlah besar atau kompleks untuk menemukan suatu pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya [10].

Data mining terbagi menjadi beberapa kelompok berdasarkan tugas yang dapat di lakukan, antara lain [11] :

1. Klasifikasi

   Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.

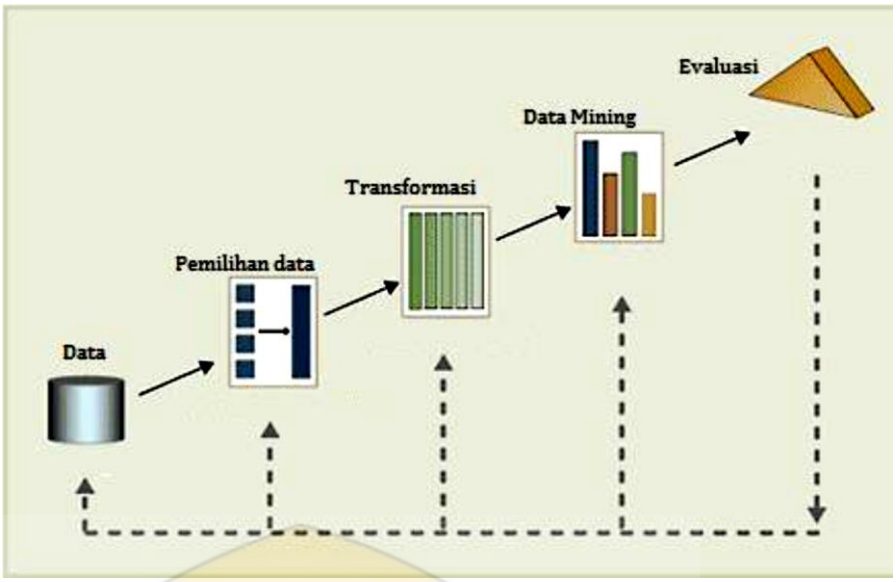
2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model di bangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, berat badan, dan level sodium darah.

3. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Contoh prediksi dalam bisnis yaitu, Prediksi harga beras dalam tiga bulan yang akan datang.

Menurut Santosa *Knowledge Discovery In Database (KDD)* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menentukan keteraturan, pola atau hubungan dalam sebuah set data yang berukuran besar [12]. *Knowledge Discovery In Database (KDD)* dapat memproses seluruh data non-trivial untuk mengetahui pola dalam data, yang mana pola yang ditemukan memiliki sifat sah dan dapat /mudah dipahami :



Gambar 2.1 Proses *KDD* (Santosa, 2007)

Adapun tahapan *Knowledge Discovery In Database (KDD)* adalah :

1. Data

   Membuat himpunan data target, penetapan himpunan data dan memfokuskan pada subset variabel atau sampel data, dimana penelitiaan akan dilakukan.

2. Pemilihan Data

   Langkah pertama pemprosesan data dan pembersihan data adalah tindakan dasar seperti penghapusan *noise*. Sebelum melakukan proses data mining, maka diperlukan proses *cleaning* pada data yang menjadi fokus dalam *Knowledge Discovery In Database*

3. Transformasi

   Pada tahap ini merupakan tahapan proses kreatif dan sangat tergantung pada pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

   Dalam pemilihan algoritma data mining untuk melakukan pencarian proses data mining yaitu antara lain teknik, metode atau algoritma dalam data mining sangat bervariasi. Penetapan metode atau algoritma yang tepat tergantung pada tujuan dan proses *KDD* secara keseluruhan.

5. Evaluasi

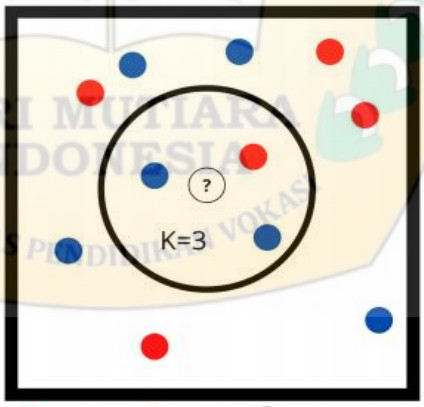
   Tahap ini merupakan tahapan pemerikasaan, apakah pola yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Fungsi data mining merupakan untuk mengklasifikasikan pola yang harus ditemukan dalam data mining. Berikut operasi-operasi dan teknik yang berhubungan dengan data mining [11]:

1. *Operasi Predictive Modeling* : (*classification, value prediction*)
2. *Database segmentation* : (*demographic clustering, neural clustering*)
3. *Link Analysis* : (*association discovery, sequential pattern discovery, similar timesequence discovery*)
4. *Deviation detection* : (*statistics, visualization)*

2.2 *k-Nearest Neighbors* (k-NN)

Algoritma *k-Nearest Neighbors* (k-NN) salah satu teknik klarisifikasi data yang kuat, dengan cara mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama berdasarkan pencocokan bobot [13]. *k-Nearest Neighbors* adalah suatu metode algoritma *supervised learning*, dimana kelas yang paling banyak muncul (mayoritas) yang akan menjadi kelas hasil klasifikasi [14]. *k-Nearest Neighbors* termasuk dalam kelompok *instance-based learning*. *k-Nearest Neighbors* merupakan contoh algoritma berbasis pembelajaran, dimana *dataset* pelatihan (training) disimpan, sehingga klasifikasi untuk *record* baru yang tidak diklasifikasi didapatkan dengan membandingkan *record* yang paling mirip dengan data latih.



Gambar 2.2 Ilustrasi KNN

Pada gambar di atas dijelaskan jika terdapat dua kelas atau label dan menggunakan nilai k=3. Terdapat 3 objek terdekat dengan objek yang belum memiliki kelas: 2 objek biru dan 1 objek merah. Maka, objek yang tidak

memiliki kelas akan memiliki kelas biru karena objek terbanyak terdekat dengan nilai *k*=3 adalah objek dengan kelas biru.

Metode *k*-NN dibagi menjadi dua fase, yaitu pembelajaran (*training*) dan klasifikasi atau pengujian (t*esting*) [13]. Secara umum untuk mendefinisikan jarak antara dua objek x dan y, digunakan rumus jarak *Euclidien* pada persamaan:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}\left(X_{training} - Y_{testing}\right)^2}$$

Keterangan :

$X_{training}$: data training ke-*i*,

$Y_{testing}$: data testing,

i : *record* (baris) ke-*i* dari tabel,

n : jumlah data training.

Dimana matriks *distance* adalah jarak skala dari kedua vektor x dan y dari matriks dengan ukuran n dimensi. Pada fase *training*, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data sample. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk *testing* data (yang klasifikasinya tidak diketahui). Jarak dari vektor baru yang ini terhadap seluruh vektor *training* sample dihitung dan sejumlah k buah yang paling dekat diambil.

2.3 Klasifikasi

Klasifikasi merupakan proses penemuan model (fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui [15]. Teknik klasifikasi yang banyak digunakan secara luas, diantaranya adalah *Neural, Rough sets, K-nearest neighbor, Bayesian classfiers network* dan lain-lain.

Dalam proses klasifikasi data terdiri dari 2 langkah, yaitu *learning* (fase training) dan klasifikasi [13]. Proses learning dibuat untuk mengalisa data training lalu direpresentasikan berupa rule klasifikasi. Sedangkan proses klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. Model tersebut dibangun dengan menganalisa database tuple. Setiap tuple diasumsikan menjadi *predefined class* yang ditentukan oleh suatu attribut yang disebut *class* label atribut. Dapat di ilustrasikan pada gambar 2.3 di bawah ini :

Gambar 2.3 Model Klasifikasi

Proses klasifikasi didasarkan pada empat komponent [16]:

1. Kelas

   Variabel dependen berupa kategori yang mempresentasikan “label” yang terdapat pada objek.

2. *Predictor*

   Variabel independen yang direpresentasikan oleh karakteristik data.

3. *Training dataset*

   Satu set data yang mempunya nilai dari kedua komponen yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor*.

4. *Testing dataset*

   Berupa data baru yang diklasifikasian oleh model data yang telah di buat dan akurasi klasifikasi di evaluasi.

2.4 *Scikit-Learn (sklearn)*

Scikit-learn atau *sklearn* adalah modul untuk bahasa pemograman python yang dibangun diatas *NumPy*, *SciPy*, dan *matplotlib*, fungsinya dapat membantu melakukan processing data ataupun melakukan training data untuk kebutuhan *machine learning* [17]. Ada banyak fitur yang dapat digunakan dengan sklearn ini, seperti C*lassification, Regression, Clustering, Dimensionality reduction, Model selection,* dan *Preprocessing data*.

Teknik *Mechine Learning* terdiri dari dua yaitu *Supervised Learning*, dan *Unsupervised Learning* [18]. Dalam *Supervised Learning* data dan label yang akan di train atau test sudah ditentukan diawal, sedangkan *Unsupervised Learning* label datanya tidak diketahui sehingga prosesnya menentukan label berdasarkan persamaan dari data-datanya.

2.5 Perangkat Lunak Yang Digunakan

Kebutuhan perangkat adalah pengumpulan kebutuhan-kebutuhan dari semua elemen sistem perangkat yang akan digunakan dalam memprediksi

kelulusan mahasiswa menggunakan *k*-NN. Adapun perangkat keras dan perangkat lunak yang digunakan antara lain adalah sebagai berikut :

a. Perangkat Keras

Perangkat keras komputer yang digunakan untuk pengklasifikasi menggunakan *k*-NN sebagai berikut :

1. ACER Aspire E 14
2. Prosessor : Intel® Core™ i3-6006U (2.0 GHz, 3MB L3 Cache)
3. Memory : 4GB DDR4
4. Storage: 500GB HDD

b. Perangkat Lunak

Perangkat lunak yang digunakan untuk pengklasifikasi menggunakan *k*-NN sebagai berikut :

1. Sistem Operasi : Windows 10
2. Jupyter Notebook