BAB II
LANDASAN TEORI

2.1 Data Mining

Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis (Hermawati, 2013). *Data mining* adalah sebuah proses untuk menemukan pola atau pengetahuan yang bermanfaat secara otomatis dari sekumpulan data yang berjumlah banyak, data mining sering dianggap sebagai bagian dari *Knowledge Discovery in Database (KDD)* yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data (Sunjana, 2010).

Suatu penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Data mining disebut juga sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data.

Menurut Sunjana (2010), Proses KDD secara garis besar dapat dijelaskan sebagai berikut.

1. *Data selection.*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/cleaning*.

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). Selain itu dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eskternal.

3. *Transformation*.

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Interpretation/evaluation*.

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.2 Naive Bayes Classifier

Naive Bayes Classifier merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (Prasetyo, 2012). Algoritma

Naive Bayes Classifier adalah salah satu algoritma yang terdapat pada teknik klasifikasi. *Naive Bayes Classifier* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Metode ini memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan *Naive* yang mengasumsikan kondisi antar atribut saling bebas. Klasifikasi *Naive* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya (Bustami, 2014). Konsep Probabilitas merupakan salah satu model statistik. Salah satu metode yang menggunakan konsep probabilistik adalah *Naive Bayes*. Algoritma *Naive Bayes* adalah salah satu algoritma dalam teknik klasifikasi yang mudah diimplementasikan dan cepat prosesnya. Pada metode ini, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain.

Dasar formula teorema Bayes yang digunakan adalah.

$$P(X|H) = \frac{P(H|X)P(H)}{P(X)} \quad \ldots(2.1)$$

Keterangan.

X = Data dengan kelas yang belum diketahui

H = Label Kelas

P(H) = Probabilitas dari Hipotesa H

P(X) = Probabilitas X

P(H|X) = Probabilitas Hipotesis H berdasarkan kondisi X.

P(X|H) = Probabilitas X, berdasarkan kondisi hipotesis H

Untuk menghitung nilai Conditionally Independent dengan rumus.

$$\rho\,(\text{н} \mid x) = \prod_{k=1}^{n} P\,(X_k \mid H) \quad \ldots.(2.2)$$

Parameter $\rho\,(\text{н} \mid x)$ adalah probabilitas vektor X pada kelas H, parameter $\prod_{k=1}^{n} P\,(X_k \mid H)$ adalah probabilitas kelas H dari semua fitur dalam vector. Selanjutnya menghitung kelas label maksimum dengan rumus.

$$\text{Probabilitas Posterior kelas} = \rho\,(x \mid \text{н}) * \rho\,(\text{н}) \quad \ldots(2.3)$$

Probabilitas class pada setiap kriteria di atas didapat dari.

$$\text{Probabilitas value}_n \text{ atribut} = \frac{\text{jumlah value}_n \text{ data pada class}}{\text{total data pada class}} \quad \ldots(2.4)$$

2.3 Data Training

Data training adalah kumpulan data masukan dan keluaran berdasarkan pengetahuan yang telah dikumpulkan sebelumnya. Dataset untuk proses *training* harus mencukupi dan mewakili setiap kondisi yang hendak diselesaikan. Terbatasnya dataset dapat menyebabkan akurasi analisis menjadi rendah dan tidak tepat (Prasetyo, 2012).

2.4 Data Uji

Data uji adalah kumpulan data masukan dan keluaran yang digunakan untuk kegiatan uji analisa metode. Data uji bersifat bebas dan sesuai dengan fakta-fakta

yang terjadi di lapangan (Prasetyo, 2012). Pengolahan data uji mengacu pada training sehingga dapat tercapai kesimpulan analisa yang diharapkan.

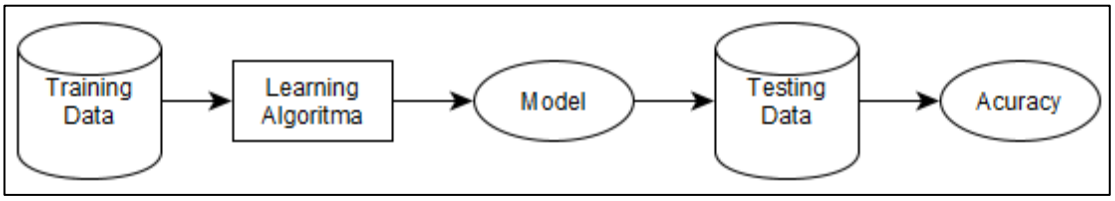
2.5 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas suatu objek yang labelnya tidak diketahui. Dalam mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data kedalam kelas – kelas yang berbeda berdasarkan aturan atau fungsi tertentu. Model itu sendiri bisa berupa aturan "jika-maka", berupa pohon keputusan, atau formula matematis.

Klasifikasi adalah menentukan sebuah *record* data baru ke salah satu dari beberapa katagori (atau kelas) yang telah didefinisikan sebelumnya (Hermawati, 2013). Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai prototipe untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya (Prasetyo, 2012).

Menurut Fajar Astuti Hermawati (2013), terdapat dua jenis model klasifikasi, yaitu.

1. Pemodelan Deskriptif (*descriptive modelling*). Model klasifikasi yang dapat berfungsi sebagai suatu alat penjelasan untuk membedakan objek – objek dalam kelas – kelas yang berbeda.
2. Pemodelan Prediktif (*predictive modelling*). Model klasifikasi yang dapat digunakan untuk memprediksi label kelas record yang tidak diketahui Penelitian yang dilakukan adalah menggunakan pemodelan prediktif, yaitu untuk mengetahui pada kelas mana suatu permasalahan yang akan diklasifikasi



Gambar 2.1 Tahap Klasifikasi (Indriani,2014)

Pada gambar 2.1 dijelaskan proses klasifikasi terdiri dari beberapa tahap yaitu.

1. *Learning (training)* adalah pembelajaran menggunakan *data training*. Untuk *Naive Bayes Classifier*, nilai probabilitas dihitung dalam proses pembelajaran.
2. *Testing* adalah pengujian model menggunakan data *testing*.
3. *Accuracy* adalah besaran error data yang sudah disebutkan sebelumnya. Terdapat 2 pilihan yang bisa diambil, yaitu membuat model lain atau menerima model tersebut- misalnya karena batasan error tersebut diterima.

2.6 Skala Likert

Skala Likert adalah suatu skala psikometrik yang umum digunakan dalam kuisioner, dan skala yang paling banyak digunakan dalam riset berupa survei

(Maryuliana, 2016). Nama skala ini diambil dari nama Rensis Likert, yang menerbitkan suatu laporan yang menjelaskan penggunaannya. Sewaktu menanggapi pertanyaan dalam skala Likert, responden menentukan tingkat persetujuan mereka terhadap suatu pertanyaan dengan memilih salah satu dari pilihan yang tersedia.