BAB III
LANDASAN TEORI

3.1. Data Mining

Data mining adalah proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar (Tan, 2006). Data mining juga dapat diartikan sebagai serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Pramudiono, 2006). Salah satu teknik yang dibuat dalam dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model (Prasetyo, 2012).

Kata mining sendiri berarti usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar. Karena itu Data Mining sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), machine learning, statistik dan database. *Data mining* menjadi alat yang semakin penting untuk mengubah data tersebut menjadi informasi. Hal ini sering digunakan dalam berbagai praktek profil, seperti pemasaran, pengawasan, penipuan deteksi dan penemuan ilmiah. Telah digunakan selama bertahun-tahun oleh bisnis, ilmuwan dan pemerintah untuk menyaring *volume* data seperti catatan perjalanan penumpang penerbangan, data sensus dan supermarket *scanner* data untuk menghasilkan laporan riset pasar.

Alasan utama untuk menggunakan data mining adalah untuk membantu dalam analisis koleksi pengamatan perilaku. Data tersebut rentan terhadap *collinearity* karena diketahui keterkaitan. Fakta yang tak terelakkan data mining adalah bahwa *subset/set* data yang dianalisis mungkin tidak mewakili seluruh domain, dan karenanya tidak boleh berisi contoh-contoh hubungan kritis tertentu dan perilaku yang ada di bagian lain dari domain . Untuk mengatasi masalah semacam ini, analisis dapat ditambah menggunakan berbasis percobaan dan pendekatan lain, seperti *Choice Modelling* untuk data yang dihasilkan manusia.

3.1.1 Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (lasrore,2009) :

1. **Deskripsi**

   Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Enkripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. **Estimasi**

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik daripada kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilau variabel prediksi.

3. **Prediksi**

   Prediksi hampir sama dengan klasifikasi dan estimasi. Kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

4. **Klasifikasi**

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

5. **Pengklusteran**

   Pengklusteran merupakan pengelompokkan record, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain.

6. **Asoisasi**

   Tugas asosiasi adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

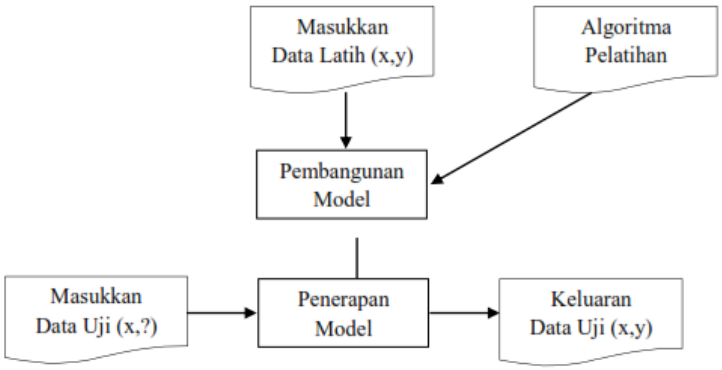
3.2. Klasifikasi

Klasifikasi merupakan pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dengan sejumlah kelas yang tesedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai protipe 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 disimpanya (Prasetyo, 2012).

Klasifikasi adalah metode data mining yang dapat digunakan untuk proses pencarian sekumpulan model (fungsi) yang dapat menjelaskan dan membedakan kelas-kelas data atau konsep, yang tujuannya supaya model tersebut dapat digunakan memprediksi objek kelas yang labelnya tidak diketahui atau dapat memprediksi kecenderungan data-data yang muncul di masa depan. Metode klasifikasi juga bertujuan untuk melakukan pemetaan data ke dalam kelas yang sudah didefinisikan sebelumnya berdasarkan pada nilai atribut data (Han dan Kamber, 2006).

3.2.1. Model Dalam Klasifikasi

Model dalam klasifikasi biasa diartikan sebagai suatu tempat untuk menerima masukan (data latih), kemudian mampu melakukan pemikiran terhadap masukan tersebut, dan memberikan jawaban sebagai keluaran dari hasil pemikiranya. Model tersebut dipakai untuk memprediksi kelas dari data uji. Proses Pekerjaan dalan klasifikasi dapat dilihat di gambar 3.1



Gambar 3.1 Proses Pekerjaan dalam klasifikasi (Prasetyo, 2012)

Dalam Pembangunan model selama proses pelatihan tersebut diperlukan suatu algoritma untuk membangunya, yang disebut algoritma pelatihan. Ada beberapa algoritma pelatihan antara lain *Naïve Bayes, Artificial Neuron Network, Support Vector Machine,* dan lain-lain.

3.3. Naïve Bayes Classifier

Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.

3.3.1. Teorema Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan *teorema Bayes* (atau aturan *Bayes*) dengan asumsi

independensi (ketidak tergantungan) yang kuat (naif). Dengan kata lain, dalam *Naïve Bayes*, model yang digunakan adalah "model fitur *independen*".

Dalam Bayes (terutama *Naïve Bayes)*, maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya, pada kasus klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat, dan menyusui. Di sini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan bertelur biasanya tidak menyusui. Dalam *Bayes*, hal tersebut tidak dipandang sehingga masing-masing fitur seolah tidak memiliki hubungan apapun.

Prediksi *Bayes* didasarkan pada *teorema Bayes* dengan formula umum sebagai berikut :

$$P(H|E) = \frac{P(E|H) \text{ x } P(H)}{P(E)} \quad \text{..... 3.1}$$

Penjelasan dari formula tersebut adalah sebagai berikut.

Parameter	Keterangan
P(H|E)	Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi.
P(E|H)	Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis H.
P(H)	Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun.
P(E)	Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis / bukti yang lain.

Ide dasar dari aturan *Bayes* adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan *Bayes* tersebut, yaitu :

1. Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Contoh, dalam suatu peramalan cuaca untuk memperkirakan terjadinya hujan, ada faktor yang mempengaruhi terjadinya hujan tersebut, yaitu mendung. Jika diterapkan dalam *Naïve Bayes* , probabilitas terjadinya hujan, jika bukti mendung sudah diamati, dinyatakan dengan :

$$\mathrm{P(Hujan|Mendung)} = \frac{\mathrm{P(Mendung|Hujan)}x\ \mathrm{P(Hujan)}}{\mathrm{P(Mendung)}}$$

P(Hujan|Mendung) adalah nilai probabilitas hipotesis hujan terjadi jika bukti mendung sudah diamati. P(Mendung|Hujan) adalah probabilitas bahwa mendung yang diamati akan mempengaruhi terjadinya hujan. P(Hujan) adalah probabilitas awal hujan tanpa memandang bukti apapun, sementara P(Mendung) adalah probabilitas terjadinya mendung.

Contoh tersebut dapat dikembangkan dengan menambahkan beberapa observasi yang lain sebagai bukti, semakin banyak bukti yang dilibatkan, semakin baik hasil prediksi yang diberikan. Namun, tentu saja bukti tersebut harus benar-benar berkaitan dan memberi pengaruh pada hipotesis. Dengan kata lain, penambahan bukti yang diamati tidak sembarangan. Misalnya, gempa bumi. Bukti gempa bumi tentu saja tidak berkaitan dengan hujan sehingga penambahan bukti gempa bumi dalam prediksi cuaca akan memberikan hasil yang salah. Walaupun ada bukti lain yang mempengaruhi cuaca seperti suhu udara, tetap saja ada nilai probabilitas P(Suhu) yang harus dinilai secara independen dalam teorema Bayes, yang sulit dilakukan karena suhu udara juga dipengaruhi oleh faktor lain seperti cuaca kemarin, mendung, polusi, dan sebagainya. Inilah sebabnya disebut *Naïve Bayes* (*Bayes Naif*).

Teorema Bayes juga bisa menangani beberapa bukti, misalnya ada E1, E2, dan E3, sehingga probabilitas akhir untuk hipotesis hujan dapat dihitung dengan cara berikut :

$$P(H|E1, E2, E3) = \frac{P(E1, E2, E3|H) \text{ x } P(H)}{P(E1,E2,E3)} \quad \text{.....3.2}$$

Karena asumsi yang digunakan untuk bukti adalah independen, bentuk diatas dapat diubah menjadi

$$P(H|E1, E2, E3) = \frac{P(E1|H) \text{ x } P(E2|H) \text{ x } P(E3|H) \text{ x } P(H)}{P(E1) \text{ x } P(E2) \text{ x } P(E3)} \quad \text{.....3.3}$$

Untuk contoh di atas, jika ditambahkan bukti suhu udara dan angin, bentuknya berubah menjadi :

$$P(\text{Hujan}|\text{Mendung}, \text{Suhu}, \text{Angin})$$
$$= \frac{P(\text{Mendung}|\text{Hujan}) \text{ x } P(\text{Suhu}|\text{Hujan}) \text{ x } P(\text{Angin}|\text{Hujan}) \text{ x } P(\text{Hujan})}{P(\text{Mendung}) \text{ x } P(\text{Suhu}) \text{ x } P(\text{Angin})}$$

3.3.2. *Naïve Bayes* untuk Klasifikasi

Kaitan antara *Naïve Bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam *teorema Bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika X adalah vektor masukan yang berisi fitur dan Y adalah label kelas, *Naïve Bayes* dituliskan dengan P(Y|X). Notasi tersebut berarti *probabilitas* label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga *probabilitas* akhir *(posterior probability)* untuk Y, sedangkan P(Y) disebut *probabilitas* awal *(prior probability)* Y.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X’ dapat

diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai P(Y'|X') yang didapat.

Formulasi *Naïve Bayes* untuk klasifikasi adalah

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \text{..... 3.4}$$

P(Y|X) adalah probabilitas data dengan vektor X pada kelas Y. P(Y) adalah probabilitas awal kelas Y. $\prod_{i-1}^{q} P(X_i|Y)$ adalah probabilitas *independen* kelas Y dari semua fitur dalam vector X. Nilai p(X) selalu tetap sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan :

$$P(X|Y = y) = \prod_{i=1}^{q} P(X_i|Y = y) \quad \text{.....3.5}$$

Setiap set fitur X = $\{X_1, X_2, X_3, \ldots, X_q\}$ terdiri atas *q* atribut (*q* dimensi).

Umumnya, Bayes mudah dihitung untuk fitur bertipe kategoris seperti pada kasus klasifikasi hewan dengan fitur "penutup kulit" dengan nilai {bulu, rambut, cangkang}, atau kasus fitur "jenis kelamin" dengan nilai {pria, wanita}. Namun untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan dalam *Naïve Bayes.* Caranya adalah :

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinyu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi Gaussian biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i \mid Y)$, sedangkan distribusi *Gaussian*

dikarakteristikkan dengan dua parameter : mean, μ, dan varian, σ^2. Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i adalah

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma\, ij}} \exp^{\frac{(X_{i-\mu_{ij}})^2}{2\sigma_{ij}^2}} \quad \text{..... 3.6}$$

Parameter μ_{ij} bisa didapat dari mean sampel X_i ($\bar{x}$) dari semua data latih yang menjadi milik kelas y_j, sedangkan σ_{ij}^2 dapat diperkirakan dari varian sampel (s^2) dari data latih. Contoh TABEL 3.1 adalah data latih untuk klasifikasi jenis hewan. Fitur yang menggunakan tipe numerik adalah berat. Mean dan varian untuk masing-masing kelas mamalia dan reptil dihitung sebagai berikut :

$$\bar{x}_{\text{mamalia}} = \frac{0.8+21+120+1.5+43+45}{6} = \frac{231.3}{6} = 38.55$$

$$\bar{x}_{\text{reptil}} = \frac{10+0.4+0.3+7}{4} = \frac{17.7}{4} = 4.425$$

$$s^2_{\text{mamalia}} = \frac{(0.8-38.55)^2+(21-38.55)^2+(120-38.55)^2}{6-1} \underline{+(1.5-38.55)^2+(43-38.55)^2+(45-38.55)^2}$$

$$s^2_{\text{mamalia}} = \frac{9801.275}{5} = 1960.255$$

$$s^2_{\text{mamalia}} = \sqrt{1960.255} = 44.275$$

$$s^2_{\text{reptil}} = \frac{(10-4.425)^2+(0.4-4.425)^2+(0.3-4.425)^2+(7-4.425)^2}{4-1}$$

$$s^2_{\text{reptil}} = \frac{70.9275}{3} = 23.6425$$

$$s^2_{\text{reptil}} = \sqrt{23.6425} = 4.8624$$

Untuk memperjelas penggunaan fitur bertipe numerik, kita akan menggunakan sebuah contoh data uji berupa hewan musang dengan nilai fitur penutup kulit = rambut, melahirkan = ya, berat = 15. Masuk ke kelas manakah hewan musang tersebut?

Tabel 3.1 Data latih klasifikasi hewan

Nama Hewan	Penutup Kulit	Melahirkan	Berat	Kelas

Ular	Sisik	Ya	10	Reptil
Tikus	Bulu	Ya	0.8	Mamalia
Kambing	Rambut	Ya	21	Mamalia
Sapi	Rambut	Ya	120	Mamalia
Kadal	Sisik	Tidak	0.4	Reptil
Kucing	Rambut	Ya	1.5	Mamalia
Bekicot	Cangkang	Tidak	0.3	Reptil
Harimau	Rambut	Ya	43	Mamalia
Rusa	Rambut	Ya	45	Mamalia
Kura-kura	Cangkang	Tidak	7	Reptil

Untuk menyelesaikannya, pertama kita harus mengetahui nilai probabilitas setiap fitur pada setiap kelasnya atau $P(X_i \mid Y_i)$, ringkasannya dapat dilihat pada TABEL 3.2. Selanjutnya, untuk data uji di atas, hitung nilai probabilitas untuk fitur dengan tipe numerik, yaitu berat.

$$P(\text{Berat} = 15 \mid \text{Mamalia}) = \frac{1}{\sqrt{2\pi}\,44.275} \exp^{-\frac{(15-38.55)^2}{2x1960.255}} = 0.0078$$

$$P(\text{Berat} = 15 \mid \text{Reptil}) = \frac{1}{\sqrt{2\pi}\,4.8624} \exp^{-\frac{(15-4.425)^2}{2x23.6425}} = 0.0077$$

Tabel 3.2 Probabilitas fitur dan kelas

Penutup Kulit		Melahirkan	
Mamalia	Reptil	Mamalia	Reptil
Sisik = 0	Sisik = 2	Ya = 6	Ya = 1
Bulu = 1	Bulu = 0	Tidak = 0	Tidak = 3
Rambut = 5	Rambut = 0		
Cangkang = 0	Cangkang = 2		
P(Kulit = Sisik | Mamalia) = 0	P(Kulit = Sisik | Reptil) = 0.5	P(Lahir = Ya | Mamalia) = 1	P(Lahir = Ya | Reptil) = 0.25
P(Kulit = Bulu | Mamalia) = 1/6	P(Kulit = Bulu| Reptil) = 0	P(Lahir = Tidak | Mamalia) = 0	P(Lahir = Tidak | Reptil) = 0.75
P(Kulit = Rambut | Mamalia) = 5/6	P(Kulit = Rambut | Reptil) = 0		
P(Kulit = Cangkang | Mamalia) = 0	P(Kulit = Cangkang | Reptil) = 0.5		

Berat		Kelas	
Mamalia	Reptil	Mamalia	Reptil
$\bar{x}_{\text{mamalia}} = 38.55$	$\bar{x}_{\text{reptil}} = 4.425$	Mamalia = 6	Reptil = 4

$s^2_{mamalia}$ = 1960.255	s^2_{reptil} = 23.6425	P(Mamalia) = 6/10 = 0.6	P(Reptil) = 4/10 = 0.4
$s^2_{mamalia}$ = 44.275	s^2_{reptil} = 4.8624		

Barulah kemudian menghitung probabilitas akhir untuk setiap kelas :

P(X | Mamalia) = P(Kulit = Rambut | Mamalia) x P(Lahir = Ya | Mamalia) x P(Berat = 15 | Mamalia) = 5/6 x 1 x 0.0078 = 0.0065

P(X | Reptil) = P(Kulit = Rambut | Reptil) x P(Lahir = Ya | Reptil) x P(Berat = 15 | Reptil) = 0 x 0.25 x 0.0077 = 0

Selanjutnya, nilai tersebut dimasukkan untuk mendapatkan probabilitas akhir.

P(Mamalia | X) = α x 0.6 x 0.0065 = 0.0039α

P(Reptil | X) = α x 0 x 0.4 = 0

α = 1/P(X)

nilainya konstan sehingga tidak perlu diketahui karena yang terbesar dari dua kelas tersebut tidak dapat dipengaruhi P(X). Karena nilai probabilitas akhir terbesar ada di kelas mamalia, data uji musang diprediksi sebagai kelas mamalia.