BAB II

LANDASAN TEORI

2.1 Pengertian Data Mining

Data mining, sering juga disebut sebagai *Knowledge Discovery In Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan (Santoso, 2007).

Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004). *Data mining* juga disebut serangkaian proses untuk menggali nilai berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007).

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data *warehouse*, atau penyimpanan infromasi lainnya. Data mining berkaitan dengan bidang ilmu-ilmu lain, seperti *database system*, data *warehousing*, *statisik*, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti *neural network*, pengenalan pola, *spatial data analysis, image database, signal processing* (Han, 2006).

Data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* ini juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga *knowledge discovery* (Tan, 2006).

2.2 Tahap – Tahap Data mining

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap yang diilustrasikan di Gambar, Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung dengan perantaraan knowledge base.

Tahap – tahap *data mining* ada 6 yaitu :

1. Pembersihan data (*data Cleaning*)

   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data yang tidak relevan. Pada umumnya data diperoleh, baik dari *database* suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa *data mining* yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performansi dari teknik *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Intregasi data (*data integraton*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu *database* atau file teks. Integrasi data dilakukan pada atribut-atribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi data (*data selection*)

   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.

4. Transformasi data (*data transformation*)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asoisasi dan clustering hanya bisa menerima input data kategorikal, Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

5. Proses *mining*

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi pola (*pattren evaluation*)

   Untuk mengidentifikasi pola-pola menarik kedalam knowledge *based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses *data mining*, mencoba metode *data mining* yang lebih sesuai, atau menerima hasil ini sebagai suatu hal yang diluar dugaan yang mungkin bermanfaat.

2.3 Klasifikasi

Klasifikasi (*Classification*) proses untuk menyatakan suatu objek ke salah satu kategori yang sudah didefinisikan sebelumnya (Bertalya, 2009). Klasifikasi (Classification) merupakan proses untuk menemukan sekumpulan model yang menjelaskan dan membedakan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah objek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih (training set). Sedangkan data uji (test set) digunakan untuk mengetahui tingkat akurasi dari model yang telah dihasilkan. Dalam proses klasifikasi pohon keputusan tradisional, fitur (atribut) dari tupel adalah kategorikal atau numerikal.

Biasanya definisi ketepan nilai (point value) sudah didefinisikan di awal. Pada banyak aplikasi nyata, terkadang muncul suatu nilai yang tidak pasti. (Tsang, Smith., 2009). Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu objek data. Metode inilah yang digunakan dalam tugas akhir ini.

2.4 Naïve Bayes Classifier

2.4.1 Teorema Bayes

Bayes merupakan teknik prediksi probalistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam *Naïve Bayes*, model yang di gunakan adalah ”model fitur independen” (Prasetyo, E. 2012).

Dalam Bayes (terutama Naïve Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidak adanya fitur lain dalam data yang sama. Contohnya, pada kasus klasifikasi hewan dengna fitur penutup kulit, melahirkan, berat dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dipastikan juga menyusui. Disini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan uang bertelur 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 :

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad \text{.....(2.1)}$$

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 hipotesis H terjadi tanpa memandang bukti apapun.

P(E) = Probabilitas awal bukti E terjadi tanpa memandang hipotesis atau bukti yang lain.

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau pristiwa (H) dapat diperkirakan pada beberapa bukti (E) yang diamati. Ada beberepa hal penting dari aturan Bayes yaitu :

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

Teorema Bayes juga bisa menangani beberapa bukti, misalnya ada E_1, E_2, dan E_3 sehinga akhir untuk hipotesis (H) dapat dihitung dengna cara berikut.

$$P(H|E_1, E_2, E_3) = \frac{P(E_1, E_2, E_3|H) * P(H)}{P(E_1, E_2, E_3)} \quad \text{.....(2.2)}$$

Karena asumsi yang digunakan untuk bukti adalah independen, bentuk di atas dapat diubah menjadi.

$$P(H| E_1, E_2, E_3) = \frac{P(E_1|H) * P(E_2|H) * |P(E_3|H) * P(H)}{P(E_1) * P(E_2) * P(E_3)} \quad \text{.....(2.3)}$$

2.5 *Naïve Bayes* untuk Klasifikasi

2.5.1 Konsep *Naïve Bayes*

Kaitan antara *Naïve bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesi 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, naive baiyes

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*). Dengan membangun model tersebut , suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai P(Y'|X') yang didapat (Prasetyo, E. 2012).

Formulasi Naïve Bayes untuk klasifikasi adalah :

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \text{.....(2.4)}$$

Keterangan :

P(Y|X) adalah probabilitas data dengan vektor X pada kelas Y. P(Y) adalah probabilitas awal kelas. $\prod_{i=1}^{q} P(X_i|Y)$ adalah probabilitas independent kelas Y dari semua fitur dalam vektor X. Karena P (X) selalu tetap, sehinga dalam perhitungan prediksi nantinya cukup hanya dengan menghitung $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang di pilih 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.

$$P(X|\ Y= y) = \prod_{i=1}^{q} P(X_i|Y = y) \quad \text{.....(2.5)}$$

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

Umumnya, metode Naïve Bayes ini mudah dihitung untuk fitur bertipe kategoris. Namun untuk tipe numerik (kontinu), ada perlakuan khusus sebelum dimasukkan dalam Naïve Bayes, yaitu :

1. Melakukan diskretisasi pada setiap fitur kontinu dan menganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinu kedalam fitur ordinal.

2. Dari distribusi probabilitas diasumsikan bentuk tertentu untuk fitur kontinu dan memperkirakan parameter distribusi dengan data peralihan. Distribusi Gaussian biasanya dipilih untuk mempresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi Gausaian dikarateristikan dengan dua parameter. mean μ, dan varian, σ^2. Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i adalah

$$P(X_i = x_i \,|Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(2.6)}$$

Keterangan :

Parameter μ_{ij} bisa di dapat 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.

2.5.2 Algoritma Klasifikasi Naïve Bayes

Algoritma Klasifikasi Naïve Bayes dihitung sesuai dengan rumus Naïve Bayes $P(Y)\prod_{i=1}^{q} P(X_i|Y)$, yang langkah-langkah perhitungannya dijelaskan sebagi berikut (Sari, M. V. 2014) :

1. Menghitung nilai probabilitas kelas berdasarkan data latih

   $\rightarrow P(Y) = \frac{Xy}{X}$

   Keterangan : X = Jumlah total data

   Xy = Nama kelas / output

2. Menghitung nilai probabilitas tiap fitur berdasarkan data latih

   $\rightarrow \prod_{i=1}^{q} P(X_i|Y) = \frac{Xyx}{\sum X\in}$

   Keterangan : Xyx = Data fitur bernilai y dari kelas X

   $\sum X \in$ = Jumlah kelas X

- Menghitung fitur bernilai numeric mengunakan rumus berikut :

$$P(X_i = x_i \mid Y = y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}},$$

Fitur numerik berikut ini dihitung tiap data uji.

3. Menghitung nilai probabilitas akhir

   Mengalikan hasil dari pobabilitas awal dan probabilitas setiap fitur atau P(Y) dan $\prod_{i=1}^{q} P(X_i \mid Y)$ pada masing-masing kelas data uji.

4. Data uji akan diklasifikasikan pada kelas dengan nilai probabilitas akhir terbesar.

2.5.3 Karateristik Naïve Bayes

Karakteristik *Naïve Bayes* (Prasetyo, E. 2012) bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas. Hal ini memberikan karakteristik *Naïve Bayes* sebagai berikut :

1. Metode *Naïve Bayes* teguh (*robust*) terhadap data-data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda (*outlier*). *Naïve Bayes* juga bisa menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan dan prediksi.
2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi *Naïve Bayes* karena asumsi independensi tersebut sudah tidak ada.

Naive Bayes memiliki beberapa keuntungan dan kekurangan yaitu sebagai berikut :

1. Keuntungan *Naive Bayes*
   - Cepat dan efisiensi ruang.
   - Kokoh terhadap atribut yang tidak relevan.

- Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi.
- Menangani Kuantitatif dan data diskrit.

2. Kekurangan *Naive Bayes*
   - Tidak berlaku jika *Probabilitas* kondisionalnya adalah nol, apabila nol maka *Probabilitas* prediksi akan bernilai nol juga.
   - Mengasumsikan Variabel bebas.

2.6 Penelitian Sebelumnya

Penelitian sebelumnya yang menggunakan metode naive baiyes adalah Kusumadewi, Sri., (2009), lulusan Universitas Islam Indonesia. Penelitiannya berjudul klasifikasi status gizi menggunakan naive baiyes classification. Atribut yang dilakukan sebagai data latih sistem adalah tinggi badan, berat badan, sex, lingkar pergelangan, lingkar perut dan status gizi. Status gizi digunakan sebagai kelas data.

Pada penelitian ini dilakukan pengukuran antropometri terhadap 47 sampel mahasiswa Teknik Informatika UII. Usia sampel berkisar antara 19 hingga 22 tahun. Ada 5 variabel pengukuran, yaitu tinggi badan (cm), berat badan (cm), jenis kelamin (cm), lingkar pergelangan (cm), dan lingkar perut (cm). lingkar pergelangan diukur dari pergelangan tangan yang tidak aktif (tangan kiri untuk status normal dan tangan kanan untuk kidal).

Data latih untuk penelitian ini adalah 47 data dan data ujinya menggunakan 47 data. Uji coba dilakukan dengan melakukan pengitungan mean dan standar deviasi setiap variabel yang bernilai kontinu (seperti tinggi badan, berat badan, lingkar pergelangan tangan dan lingkar perut), kemudian menghitung probabilitas untuk setiap kategori itu sendiri dan setelah itu menghitung probabilitas setiap kategori apabila diberikan input tertentu. berdasarkan hasil pengujian terhadap semua data, diperoleh hasil bahwa terdapat 38 yang tidak sesuai kelas yang diberikan dan 9 hasil yang tidak sesuai dengan hasil yang diberikan.

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma naive baiyes dapat digunakan sebagai salah satu metode untuk klasifikasi status gizi berdasarkan hasil pengukuran antroprometri dan model sistem yang dibangun memiliki kinerja yang baik karena hasil pengujian menunjukkan total kinerja sebesar 0,932 atau 93,2 %.

Syawli, Almira., dkk (2012), lulusan Universitas Brawijaya Malang. Penelitian ini berjudul Diagnosa penyakit diabetes mellitus dengan metode *Naïve Bayes* berbasis desktop application. Atribut yang digunakan ada 22 yaitu, polituria, polidipsia, polipagia, kesemutan, rasa tebal, berat badan turun, kulit, gatal, bisul, infeksi, keputihan, luka, lapar, gemetar, lemah, konsentrasi, keringat, berdebar, pusing, gelisah dan koma. Dalam penelitian ini menggunakan 18 data set dimana data latih 10 dan data uji 8.

Langkah pertama yaitu menghitung probabilitas kemunculan dari setiap fitur (gejala) terhadap kelasnya. Kemudian melakukan pengujian keakuratan dilakukan dengan memasukkan masalah yang sama dengan data training sebelumnya. Setelah itu dilakukan analisis keakuratan data dengan memasukkan data sample melalui sistem. Dari pengujian tersebut, dapat diketahui tingkat keakurasiannya yaitu 94,4 % dan errornya adalah 5,6 % .

Nurvenus, Karid., dkk (2012), lulusan Universitas Brawijaya Malang. Penelitian ini berjudul Klasifikasi bawang merah, putih dan bombay menggunakan metode *naïve bayesian classifier.* Atribut yang digunakan ada 3 yaitu *Red, Green* dan *Blue*. Kemudian menguji jenis bawang berdasarkan warna RGB dan diklasifikasikan kedalam 3 kelas yaitu kelas bawang merah, bawang putih dan bawang Bombay. Metode yang digunakan adalah metode *Naïve Bayes* dengan 75 data set yang akan dilatih. Ada beberapa faktor yang menyebabkan suatu bawang tidak diklasifikasikan oleh program sesuai jenisnya. Salah satunya adalah bawang-bawang yang tidak diklasifikasikan sesuai jenisnya tersebut disebabkan oleh faktor pencahayaan pada gambar yang membuat RGB dari bawang yang diamati berbeda. Dengan 75 data latih dan berdasarkan pengujian, menunjukkan bahwa metode Naïve Bayes mempunyai keakuratan 80% dan error

20% dalam menentukan kelas bawang merah, kelas bawang putih atau kelas bawang Bombay.

Rahmawati, Syahriyatur., (2010) lulusan Universitas Muhammadiyah Gresik, Penelitiannya berjudul penerapan metode Fuzzy C-means untuk pengelompokan keluarga penerima beras miskin dikelurahan pekauman Kec. Gresik. Atribut yang digunakan dalam penelitian ini adalah pekerjaan, pendidikan, jumlah keluarga dan indikator (kesehatan, pendidikan, sosial budaya dll). Dari 20 data uji yang digunakan didapatkan hasil label keluarga prasejahtera sebanyak 7, label keluarga sejahtera 1 sebanyak 3, label keluarga sejahtera 2 sebanyak 5, label keluarga sejahtera 3 sebanyak 3, label keluarga sejahtera 3+ sebanyak 2.