BAB II

LANDASAN TEORI

2.1 Data Mining

Banyak orang menggunakan istilah *data mining* dan *knowledge discovery in databases* (KDD) untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu kumpulan data yang besar. Tetapi kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Salah satu tahapan dalam proses KDD adalah *data mining*.

Sedangkan Pejic bach berpendapat bahwa *data mining* adalah istilah yang menggambarkan berbagai teknik yang digunakan dalam *domain machine learning*, analisis statistik, teknik pemodelan, dan teknologi basis data yang dapat digunakan di berbagai industri. Dengan kombinasi teknik-teknik ini, dimungkinkan untuk menemukan berbagai jenis struktur dan hubungan dalam data, serta untuk mendapatkan aturan dan model yang memungkinkan prediksi dan pengambilan keputusan dalam situasi baru.

Dengan *data mining* maka akan didapat suatu pola yang nantinya dapat menjadi *knowledge* yang bermanfaat. *Data mining* sekarang sudah dipergunakan untuk mengambil *knowledge* dari basis data (*database*) organisasi.

2.2 Customer Churn

Seperti dalam industri jasa, di rumah sakit pun sering terjadi pasien yang berpindah ke rumah sakit lain, dikatakan sebagai *Churn*. Hadden, Tiwaria, Roy dan Ruta berpendapat bahwa manajemen *churn* adalah istilah yang telah diadopsi untuk mendefinisikan berpindahnya pelanggan. Lebih khusus lagi, manajemen *churn* adalah konsep mengidentifikasi pelanggan yang berniat untuk berpindah kepada pesaing. Setelah diidentifikasi, pelanggan ini dapat ditargetkan dengan kampanye pemasaran proaktif untuk upaya retensi.

2.4 Klasifikasi

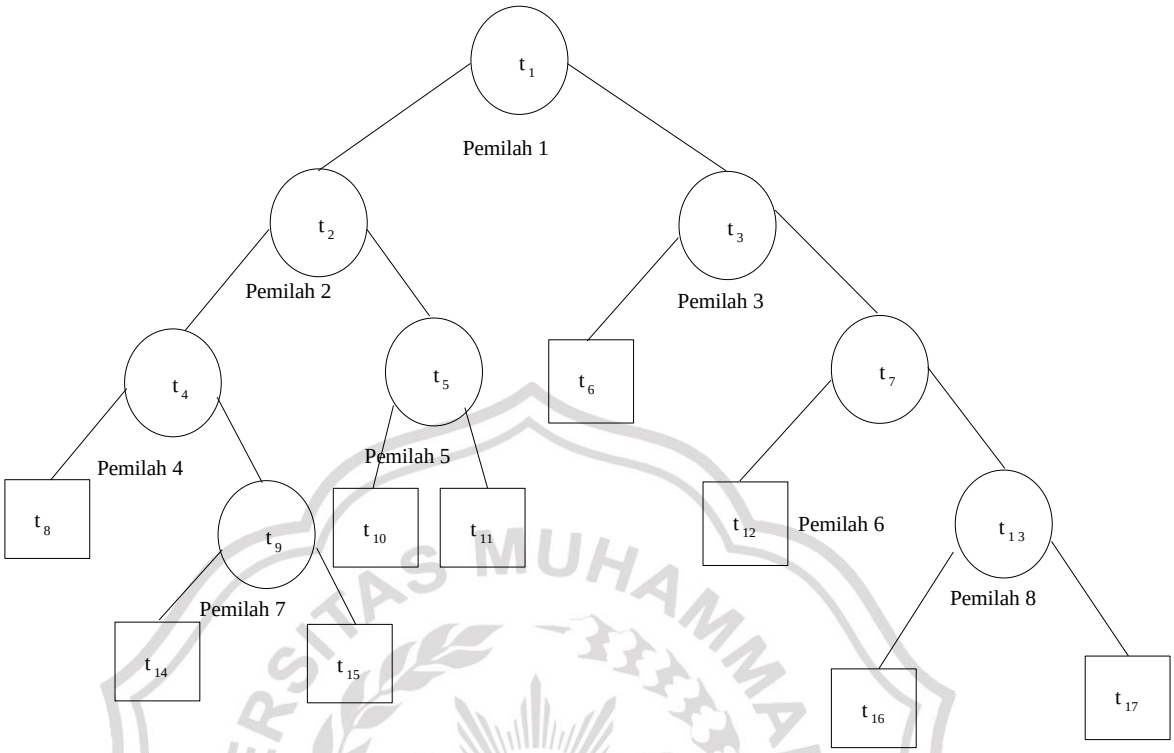
Algoritma Klasifikasi merupakan algoritma yang digunakan untuk mengklasifikasikan sesuatu berdasarkan variabel-variabel yang sudah ditentukan sebelumnya. Diperlukan dua jenis variabel yaitu variabel prediktor untuk menentukan variabel tujuan dan variabel tujuan yang ditentukan berdasarkan variabel prediktor. (Susanto, 2010).

Pohon keputusan adalah alat yang kuat dan populer untuk klasifikasi dan prediksi. Mereka menarik karena mereka mudah dipahami, karena mereka dapat disajikan secara grafis sebagai pohon juga dalam bentuk aturan (dalam bahasa Inggris atau dalam SQL). Algoritma yang paling populer adalah CHAID (Chisquared automatic induksi), CART dan C4.5 (versi terbaru dari algoritma ID3).

2.5 *Classification and Regression Trees* (CART)

Clasification and Regression Trees (CART) merupakan salah satu metode atau algoritma dari teknik pohon keputusan (*decision tree*). Metode yang dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone ini merupakan teknik klasifikasi dengan menggunakan algoritma penyekatan rekursif secara biner (*binary recursive partitioning*) (Roger dan Lewis, 2000).

Istilah "*binary*" berarti pemilahan dilakukan pada sekelompok data yang terkumpul dalam suatu ruang yang disebut simpul (*node*) menjadi dua kelompok yang disebut simpul anak (*child nodes*). Istilah "*recursive*" berarti prosedur penyekatan secara biner dilakukan secara berulangulang. Setiap simpul anak yang diperoleh dari penyekatan simpul awal kemudian bisa dipilah kembali menjadi dua simpul anak lagi, dan begitu seterusnya hingga memenuhi kriteria tertentu. Sedangkan istilah "*partitioning*" memiliki arti bahwa proses klasifikasi dilakukan dengan cara memilah suatu kumpulan data menjadi beberapa bagian atau partisi. Contoh struktur pohon klasifikasi dapat dilihat pada **gambar 2.1**.



Gambar 2.1 Struktur Pohon Klasifikasi

(Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. 1993)

Ilustrasi dari struktur pohon klasifikasi ditunjukkan pada Gambar 2.1. Simpul awal yang merupakan variabel terpenting dalam kelas amatan disebut *parent node* dengan notasi simpul t_1, simpul dalam (*internal nodes*) dinotasikan dengan t_2, t_3, t_4, t_7, t_9 dan t_{10}, serta simpul akhir (*terminal nodes*) yang dinotasikan dengan t_5, t_6, t_8, t_{11}, t_{12}, t_{13}, t_{14} dan t_{15} dimana setelahnya tidak ada lagi pemilahan. Penghitungan *depth* (kedalaman) pohon dimulai dari simpul utama t_1 yang berada pada kedalaman 1, sedangkan t_2 dan t_3 berada pada kedalaman 2 begitu seterusnya hingga t_{12}, t_{13}, t_{14} dan t_{15} yang berada pada kedalaman 5. Selain itu, setiap simpul terminal diberi tanda dengan label kelas.

Menurut Breiman ,CART akan menghasilkan pohon klasifikasi jika variabel respon mempunyai skala kategorik dan akan menghasilkan pohon regresi jika variabel respon berupa data kontinu. Tujuan utama CART adalah untuk

mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian.

Metode pengklasifikasian CART memiliki beberapa kelebihan. Pertama, CART merupakan metode nonparametrik sehingga tidak ada asumsi distribusi variabel prediktor yang perlu dipenuhi. Kedua, CART tidak hanya memberikan klasifikasi, namun juga estimasi probabilitas kesalahan pengkalsifikasian. Ketiga, metode ini memudahkan dalam hal eksplorasi dan pengambilan keputusan pada struktur data yang kompleks dan multivariabel karena struktur data dapat dilihat secara visual. Keempat, hasil klasifikasi akhir berbentuk sederhana dan mengklasifikasikan data baru secara efisien. Kelima, kemudahan dalam menginterpretasi hasil.

2.5.1 Pembentukan Pohon Klasifikasi

Pembentukan pohon klasifikasi diawali dengan menentukan variabel dan nilai dari variabel tersebut (*threshold*) untuk dijadikan pemilah tiap simpul. Dalam prosesnya, pembentukan pohon klasifikasi dibutuhkan data *learning* sampel L yang terdiri atas N pengamatan. Menurut Breiman, et al (1993), proses pembentukan pohon klasifikasi terdiri atas 3 tahapan yaitu sebagai berikut.

a. Pemilihan Pemilah

Pada tahap ini, data yang digunakan adalah sampel data *learning* L yang kemudian dipilah berdasarkan aturan pemilahan dan kriteria *goodness of split*. Himpunan bagian yang dihasilkan dari proses pemilahan harus lebih homogen dibandingkan pemilahan sebelumnya. Hal ini dilakukan dengan mendefinisikan keheterogenan simpul (*impurity* atau *i(t)*). Menurut Breiman, et al (1993), fungsi keheterogenan yang sangat mudah dan sesuai diterapkan dalam berbagai kasus adalah Indeks Gini. Indeks Gini akan selalu memisahkan kelas dengan anggota paling besar atau kelas terpenting dalam simpul tersebut terlebih dahulu. Pemilahan yang memberikan

nilai penurunan keheterogenan tertinggi merupakan pemilahan terbaik. Fungsi Indeks Gini dituliskan dalam persamaan berikut :

$$i(t) = \sum i, j = 1 p(i|t) p(j|t) \ldots\ldots\ldots\ldots (2.1)$$

Dengan *p(j|t)* adalah proporsi kelas *j* pada simpul *t* dan *p(i|t)* adalah proporsi kelas *i* pada simpul *t*.

Pemilahan simpul dimulai dengan memeriksa nilai–nilai variabel independen dan dilakukan secara rekursif pada setiap simpul dengan dua tahapan. Tahapan yang pertama adalah mencari semua kemungkinan pemilah pada variabel prediktor. Menurut Breiman, et al (1993), proses pemilahan simpul menjadi dua simpul anak dilakukan dengan mengikuti aturan sebagai berikut.

1. Setiap pemilahan hanya bergantung pada nilai yang berasal dari satu variabel prediktor saja.
2. Apabila variabel prediktor berskala kontinu, maka pemilahan yang diperbolehkan adalah $x_j \leq c_i$ dan $x_j > c_i$, dengan $i = 1, 2, \ldots, n\text{-}1$dan c_i adalah nilai tengah dari dua nilai amatan sampel berurutan yang berbeda dari variabel X_j. jika suatu ruang sampel berukuran *n* dan terdapat *n* nilai amatan sampel yang berbeda pada variabel X_j, maka akan terdapat sebanyak *n*-1 kemungkinan pemilahan yang berbeda.
3. Apabila variabel prediktor berskala kategorik, maka pemilahan berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua simpul yang saling lepas (*disjoint*). Apabila variabel prediktor berskala nominal bertaraf *L*, maka akan diperoleh sebanyak 2^{L-1}-1 pemilahan yang mungkin. Akan tetapi, apabila kategori variabel prediktor berskala ordinal bertaraf *L*, maka akan diperoleh sebanyak *L*-1 pemilahan yang mungkin.

Pemilahan yang terpilih akan membentuk suatu himpunan kelas yang disebut sebagai simpul. Simpul tersebut akan melakukan pemilahan secara rekursif sampai diperoleh simpul akhir (*terminal nodes*).

Setelah dilakukan pemilahan dari semua kemungkinan pemilah, maka tahapan berikutnya adalah menentukan kriteria *goodness of split* ($\phi(s,t)$) untuk mengevaluasi pemilah dari pemilah *s* pada simpul *t*. *Goodness of split* ($\phi(s,t)$) merupakan penurunan heterogenitas, yaitu.

$$\big(\phi(s,t)\big) = \Delta i(s,t) = i(t) - P_L i(i_L) - P_R i(i_R) \ldots\ldots\ldots (2.2)$$

Dengan

$i(t)$: fungsi heterogenitas pada simpul *t*.

p_L : proporsi pengamatan menuju simpul kiri.

p_R : proporsi pengamatan menuju simpul kanan.

$i(t_L)$: fungsi heterogenitas pada simpul anak kiri.

$i(t_R)$: fungsi heterogenitas pada simpul anak kanan.

Pemilah yang menghasilkan $\phi(s,t)$ lebih tinggi merupakan pemilah terbaik karena mampu mereduksi heterogenitas lebih tinggi. Pengembangan pohon ini dilakukan dengan pencarian pemilah yang mungkin pada simpul t_1 yang kemudian akan dipilah menjadi t_2 dan t_3 oleh pemilah *s*, dan seterusnya. t_L dan t_R merupakan partisi dari simpul *t* menjadi dua himpunan bagian saling lepas dimana p_L dan p_R adalah proporsi masing-masing peluang simpul. Karena $t_L \cup t_R = t$ maka nilai $\Delta i(s,t)$ merepresentasikan perubahan dari kehetoregenan dalam simpul *t* yang semata-semata disebabkan oleh pemilah *s*. jika simpul yang diperoleh merupakan kelas yang tidak homogen, prosedur yang sama akan diulangi.

b. Penentuan Simpul Terminal

Suatu simpul *t* akan menjadi simpul terminal atau tidak, akan dipilah kembali apabila pada simpul *t* tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum sebesar *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak. Menurut Breiman, et al (1993), pengembangan pohon akan berhenti apabila pada simpul terdapat pengamatan berjumlah kurang dari atau sama dengan 5 ($n \leq 5$). Selain itu, proses pembentukan pohon juga akan berhenti apabila sudah mencapai batasan jumlah level yang telah ditentukan atau tingkat kedalaman (*depth*) dalam pohon maksimal.

c. Penandaan Label Kelas

Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p(j_0|t) = max_j p(j|t) = max_j \frac{N_j(t)}{N(t)} \text{ .....(2.3)}$$

dengan:

$p(j|t)$: proporsi kelas *j* pada simpul *t*

$N_j(t)$: jumlah pengamatan kelas *j* pada *terminal node t*

$N(t)$: jumlah total pengamatan pada *terminal node t*

Label kelas untuk simpul terminal *t* adalah j_0 yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul *t* yang paling kecil sebesar $r(t) = 1 - max_j\, p(j|t)$.