BAB II

LANDASAN TEORI

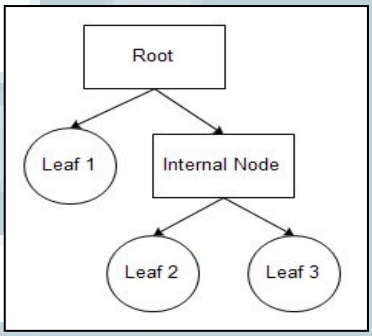
2.1 Data Mining

Data mining merupakan proses mengolah data yang terbatas dengan kemungkinan model yang tidak terbatas dan bertujuan untuk menghasilkan model yang paling baik menjelaskan data yang ada, dengan cara mengaplikasikan Algoritma data analisis dan data *discovery* (Ozer, 2008). *Data mining* juga dapat dikatakan sebagai proses mengekstrak pengetahuan dari data yang banyak (Han dan Michelin, 2006). Pengetahuan ini berupa keteraturan, pola, dan hubungan dalam *set* data yang berukuran besar dan tidak diketahui sebelumnya. Terdapat beberapa metode untuk pengolahan dalam *data mining* dan dapat dibagi menjadi 2 jenis secara umum yaitu *predictive method* dan *descriptive method* (Sondwale, 2015). *Predictive method* mengambil kesimpulan dari data yang ada untuk membuat prediksi pada data selanjutnya, klasifikasi, regresi dan deviasi merupakan beberapa contoh teknik pada *predictive method*. *Descriptive method* mengeneralisasikan karakterisitik data yang terdapat dalam database, *clustering, association,* dan *sequential mining* merupakan beberapa contoh teknik pada *descriptive method*.

2.2 Decision Tree/ Classification Tree

Pembuatan model untuk memprediksikan kelas suatu objek berdasarkan atribut yang dimilikinya merupakan hal yang harus dilakukan dalam *data mining* dengan *predictive method*. *Decision tree* merupakan salah satu metode untuk melakukan pembuatan model tersebut dan cukup terkenal karena mudah untuk

diinterprestasikan, tingkat akurasi yang baik, dan efisien dalam menangani atribut diskret ataupun numerik/kontinyu (Bening, 2014). *Decision tree* mempunyai konsep mengubah data menjadi pohon keputusan lalu pohon keputusan diubah menjadi aturan-aturan keputusan. Data yang terdapat pada pohon keputusan biasanya berbentuk tabel yang memiliki atribut dan *record.* Atribut menyatakan parameter yang digunakan sebagai kriteria dalam pembentukan tree.



Gambar 2.1 Decision Tree

Decision tree terdiri dari Simpul akar, yang tidak memiliki cabang masukan dan berpengaruh paling besar pada suatu kelas tertentu. Simpul internal yang merepresentasikan test atau subset dari sebuah Atribute dan simpul daun yang merupakan sebuah sambungan *node* dari *tree* berupa *class label*.

2.3 Algoritma C4.5

Algoritma C4.5 merupakan salah satu Algoritma klasifikasi dalam data mining, yang menggunakan model *decision tree*. Algoritma C4.5 merupakan penerus ID3 (*Iterative Dichotomiser*) yang mengadopsi greedy/*nonbacktracking* dimana *decision tree* dibangun dengan cara atas kebawah (*top down*), diulang (*recursive*), dan dibagi lalu diselesaikan (*divide and conquer*) (Han dan Michelin, 2006). Algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan Algoritma ID3, tetapi memiliki beberapa perbedaan yaitu : Dapat menangani

atribut diskrit dan numerik/kontinyu, dapat menangani training data dengan *missing value*, hasil pohon keputusan dipangkas (*pruning*) setelah dibentuk, pemilihan atribut dilakukan menggunakan *gain ratio.* Algoritma C4.5 menggunakan metode *pessimistic pruning*, metode *pessimistic pruning* memangkas *subtree* didasarkan dari tingkat kesalahan *training dataset* dan tidak memerlukan *prune set*. Dalam memilih kelas *default*, C4.5 memilih kelas yang memiliki *training tuples* dengan aturan yang paling sedikit.

Terdapat beberapa tahapan dalam membangun *decision tree* menggunakan Algoritma C4.5 yaitu :

1. Pilih Atribut sebagai akar berdasarkan *gain* paling tinggi
2. Nilai pada Atribut yang terpilih akan menjadi cabang.
3. Tentukan Atribut selanjutnya untuk menjadi cabang berdasarkan klasifikasi nilai dari Atribut sebelumnya menggunakan *gain* nilai tersebut.
4. Ulangi langkah pertama sampai langkah ketiga sampai semua syarat terpenuhi.
5. Syarat tersebut adalah :
    - Semua kasus pada cabang memiliki kelas yang sama
    - Tidak ada Atribut yang dapat dipartisi

Atribut yang dipilih sebagai akar adalah atribut yang memiliki nilai *gain* paling tinggi. Atribut numerik akan menggunakan *gain ratio*, sedangkan atribut diskrit menggunakan *information gain*. Nilai *gain ratio* didapat dari hasil pembagian *information gain* dengan *split info*. Nilai *information gain* didapat dari nilai entropy target dikurang nilai entropy tiap *values/record* dalam atribut. Entropi merupakan peluang kemunculan suatu *record* dalam atribut.

Rumus *information gain* adalah sebagai berikut :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i)$$ ..Rumus 2.1

Keterangan :

S : Himpunan Kasus

A : Atribut

n : Jumlah partisi Atribut A

|Si| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

Rumus Entropy target adalah sebagai berikut :

$$Entropy(S) = \sum_{i=1}^{n} - p_i \times \log_2 p_i$$ ..Rumus 2.2

Keterangan :

S : Himpunan kasus

n : Jumlah partisi S

pi : Proporsi dari Si terhadap S

Rumus Gain Ratio adalah sebagai berikut :

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$ ..Rumus 2.3

Keterangan :

S = ruang (data) sample yang digunakan untuk training

A = atribut

Rumus Split Information adalah sebagai berikut :

$$SpliInfo(S,A) = - \sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$ ..Rumus 2.4

Keterangan :

S = ruang (data) sample yang digunakan untuk training

A = atribut

S_i = jumlah sampel untuk atribut atribut i

Terdapat dua kunci dalam pencarian entropi yaitu :

- Jika sample yang diberikan memiliki nilai yang sama/*homogeneous* maka akan memiliki entropi 0 (nol)
- Jika sampel yang diberikan memiliki nilai yang terbagi sama rata/*equally divided* maka akan memiliki entropi 1 (satu)

2.4 Pinjaman KPR

Peminjaman KPR (kredit peminjaman rumah) merupakan salah satu pelayanan yang diberikan oleh bank khusus untuk pembangunan atau renovasi rumah (Fatmasari, 2013). Secara umum KPR dibagi menjadi dua jenis yaitu KPR subsidi dan KPR non Subsidi. Perbedaan utama jenis KPR ini adalah suku bunga yang rendah dan cicilan ringan tanpa ada perubahan bunga sepanjang jangka waktu kredit. KPR subsidi diawasi oleh pemerintah dan memiliki syarat-syarat khusus yang harus dipenuhi seperti penghasilan pemohon dan maksimum kredit yang diberikan. Syarat-syarat tersebut adalah sebagai berikut : warga negara indonesia, telah berusia 21 tahun atau telah menikah, belum pernah memiliki hunian, belum pernah menerima subsidi perumahan, termasuk dalam kategori masyarakat berpenghasilan rendah, yang memiliki pekerjaan atau penghasilan tetap, dan memiliki NPWP dan SPT tahunan.

2.5 K-Fold Cross Validation

Cross-validation merupakan metode untuk mengevaluasi dan membandingkan Algoritma pembelajaran dengan cara membagi data menjadi dua bagian, satu bagian digunakan untuk *learning* dan *testing*, satu bagian lagi

digunakan untuk memvalidasi model Algoritma yang digunakan (Refaeilzadeh, 2009). *K-fold cross validation* merupakan bentuk dasar dari metode *cross validation*. Pada *K-fold cross validation* data dibagi oleh k agar terbagi menjadi segmen-segmen yang merata dan data diuji sebanyak k kali, dimana setiap kali pengujian k sisa data menjadi data *learning*. Dalam data *mining* dan *machine learning* data cenderung dibagi menjadi sepuluh segmen (Refaeilzadeh, dkk, 2009). *Cross-validation* mempunyai dua tujuan utama yaitu menghitung estimasi performa dari mode Algoritma *learning*, dan membandingkan dua atau lebih Algoritma *learning*.

Cross validation dapat memberikan indikasi seberapa baik prediksi sebuah Algoritma pembelajaran jika diberikan data *set* yang belum pernah dipelajari sebelumnya. *Cross validation* dapat memberikan indikasi dengan cara yang sudah dijelaskan yaitu mengambil beberapa data sebelum dilakukan pembelajaran. Setelah dilakukan pembelajaran, data yang sudah diambil dapat diuji untuk lalu dilakukan validasi silang dengan hasil yang sudah diketahui. *Cross validation* memiliki beberapa jenis mulai dari *holdout* yang paling dasar, *k-fold* yang membantu mengurangi variasi data dari evaluasi yang dilakukan *holdout*, sampai *leave–one-out* yang merupakan jenis paling kompleks.

2.6 Prinsip 5C

Pemohon kredit akan dianalisis berdasarkan *character*, *capital*, *capacity*, *collateral*, dan *constraint* (Wardiah, 2013). *Character* adalah keadaan waktu atau sifat nasabah. *Capital* jumlah dana atau modal sendiri. *Capacity* kemampuan menjalankan usaha/pekerjaan. *Collateral* merupakan barang yang diserahkan sebagai agunan. *Constraints* adalah batas pelaksaan bisnis.