ISSN: 1907-5022

KLASIFIKASI DATA NASABAH SEBUAH ASURANSI MENGGUNAKAN ALGORITMA C4.5

Sunjana
Universitas Widyatama
E-mail: sunjana@widyatama.ac.id

ABSTRAKS

Pada penelitian ini, penulis berusaha menambang data(data mining) nasabah sebuah perusahaan asuransi untuk mengetahui lancar atau tidak lancarnya nasabah tersebut. Data yang ada dianalisis menggunakan pendekatan pohon keputusan(decision tree) yaitu algoritma C4.5. Dengan algoritma ini dapat diketahui data nasabah mana yang dikelompokkan ke kelas lancar dan data nasabah mana yang dikelompokkan ke kelas tidak lancar. Kemudian pola tersbut dapat digunakan untuk memperkirakan nasabah yang bergabung, sehingga perusahaan bias mengambil kepusan menerima atau menolak calon nasabah tersebut.

Kata Kunci: data mining, pohon keputusan,algoritma c4.5.

1. PENDAHULUAN

Kehadiran data mining dilatarbelakangi dengan masalah *data explosion* yang dialami akhir-akhir ini dimana banyak perusahaan / bank / organisasi telah mengumpulkan data sekian tahun lamanya (data pembelian, data penjualan, data nasabah, data transaksi, dan lain-lain). Seperti yang terjadi pada sebuah perusahaan asuransi, data yang tersimpan selama ini hanya sebagai dokumentasi dan hanya dipakai untuk kebutuhan transaksi saja. Pertanyaannya sekarang, apakah data tersebut akan dibiarkan menggunung, tidak berguna lalu dibuang, ataukah kita dapat me-"nambang"-nya untuk mencari 'emas' dan 'berlian' yaitu informasi yang berguna untuk organisasi kita.

Pertumbuhan yang pesat dari akumulasi data itu ternyata menciptakan kondisi yang sering disebut sebagai *"Rich of Data but Poor of Information"* karena data yang terkumpul itu tidak dapat digunakan untuk pengambilan keputusan. Kumpulan data itu dibiarkan begitu saja seakan-akan menjadi sebuah "kuburan data *(data tombs)"*.

Kebutuhan dari dunia bisnis yang ingin memperoleh nilai tambah dari data yang telah mereka kumpulkan telah mendorong penerapan teknik - teknik analisa data dari berbagai bidang seperti statistik, kecerdasan buatan, *database* dan lain sebagainya pada data berskala besar itu yang akhirnya memunculkan metodologi baru yang disebut *data mining*.

Pada penelitian ini, penulis bermaksud untuk menerapkan ***Metode Algoritma C4.5***"., dalam menambang data nasabah, sehingga dapat dicari pola status nasabah untuk dapat dijadikan bahan analisis perusahan dalam menentukan calon nasabah dimasa yang akan datang.

2. LANDASAN TEORI

Algoritma C4.5. merupakan kelompok algoritma decision tree. Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* berupa data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenaranya. Sedangkan *samples* merupakan *field-field* data yang nantinya akan kita gunakan sebagai parameter dalam melakukan klasifikasi data. Berikut adalah algoritma dasar C4.5

Algoritma C4.5

(1) *Build the decision tree form the training set (conventional ID3).*
(2) *Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to a leaf node.*
(3) *Prune (generalize) each rule by removing preconditions that increase classification accuracy.*
(4) *Sort pruned rules by their accuracy, and use them in this order when classifying future test examples.*

3. INFORMATION GAIN

Information gain adalah salah satu a*tribute selection measure* yang digunakan untuk memilih test atribut tiap node pada *tree*. Atribut dengan *information gain* tertinggi dipilih sebagai test atribut dari suatu node [6].

Ada 2 kasus berbeda pada saat penghitungan *information gain*. Pertama untuk kasus penghitungan atribut tanpa *missing value* dan kedua, penghitungan atribut dengan *missing value*.

3.1 Penghitungan Information Gain tanpa Missing Value

Misalkan S berisi *s* data samples. Anggap atribut untuk class memiliki *m* nilai yang berbeda,

ISSN: 1907-5022

C_i (untuk i = 1, …,I). anggap s_i menjadi jumlah samples S pada class C_i. Maka besar information-nya dapat dihitung dengan :

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$$

Dimana $p_i = \frac{si}{s}$ adalah probabilitas dari sample yang mempunyai class C_i.

Misalkan atribut A mempunyai *v* nilai yang berbeda, $\{a_1, a_2, ..., a_v\}$. Atribut A dapat digunakan untuk mempartisi S menjadi *v* subset, $\{S_1, S_2, ..., S_v\}$, dimana S_j berisi samples pada S yang mempunyai nilai a_j dari A. Jika A terpilih menjadi test atribut (yaitu, best atribut untuk splitting), maka subset-subset akan berhubungan dengan pertumbuhan node-node cabang yang berisi S. Anggap s_{ij} sebagai jumlah samples class C*i* pada subset S_j. Entropy, atau nilai information dari subset A adalah :

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_1, s_2, \ldots, s_m)$$

$\frac{s_{1j} + \ldots + s_{mj}}{s}$ adalah bobot dari subset *j*th dan jumlah samples pada subset (yang mempunyai nilai a_j dari A) dibagi dengan jumlah total samples pada S. Untuk subset S_j, $I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} * \log_2(p_{ij})$

Dimana $p_{ij} = \frac{s_{ij}}{|s_j|}$ adalah probabilitas sample S_j yang mempunyai class C*i*.

Maka nilai information gain atribut A pada subset S adalah

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

3.2 Penghitungan Information Gain dengan Missing Value

Untuk atribut dengan *missing value* penghitungan *information gain*-nya diselesaikan dengan *Gain Ratio*. Sebelum menghitung *Gain Ratio* terlebih dahulu dihitung $I(s_1, s_2, \ldots, s_m)$ dan E(A).

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_1, s_2, \ldots, s_m)$$

Dimana penghitungan $I(s_1, s_2, \ldots, s_m)$ dan E(A) hanya dilakukan pada atribut yang ada nilainya. Kemudian untuk mencari *gain* dari atribut A dihitung dengan rumus sebagai berikut :

Gain (A) = Prob S yang diketahui * E(A)

Dimana,

A = atribut dengan *missing value* yang sedang dicari nilai *gain*-nya,

S = jumlah *samples* pada subset A yang diketahui nilainya.

Sedangkan nilai split pada atribut A dinyatakan dengan :

$$Split(A) = -u * \log_2 u - \sum_{i=1}^{m} p_j * \log_2(p_j)$$

Dimana,

u adalah prob samples pada atribut A yang merupakan *missing values*.

$p_j = \frac{s_j}{|s|}$ adalah probabilitas sample S_j yang diketahui nilainya.

Nilai *Gain Ratio* pada atribut A :

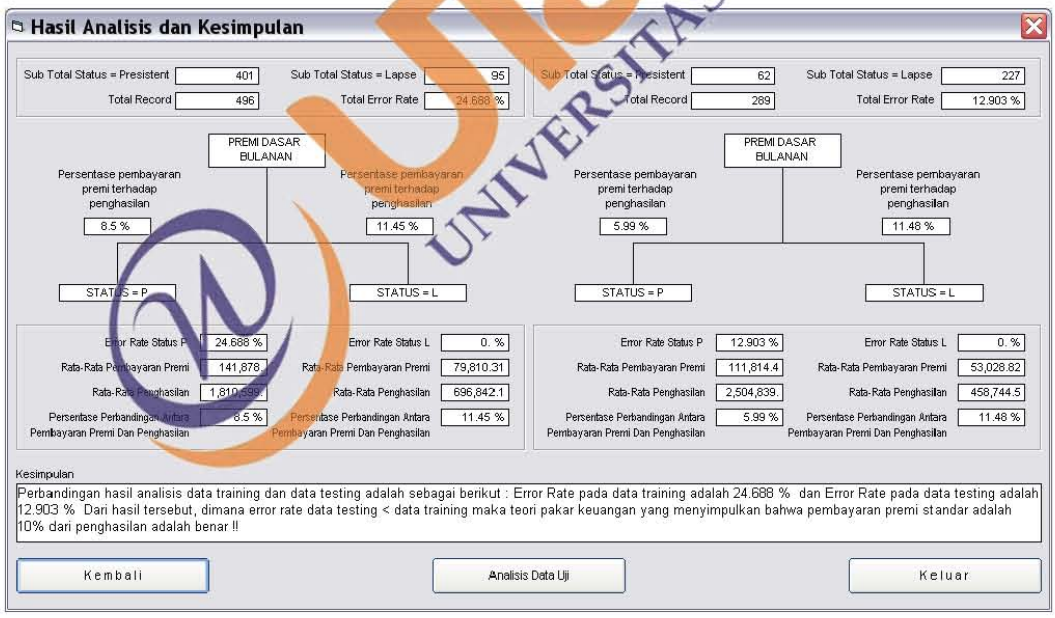
Gain Ratio(A) = Gain(A) / Split(A)

4. HASIL PENELITIAN

Atribut-atribut yang digunakan dalam penelitian ini adalah penghasilan, premi_dasar, cara_pembayaran, mata_uang, dan status sedang label yang digunakan untuk pengklasifikasian adalah lancar dan tidak lancar.

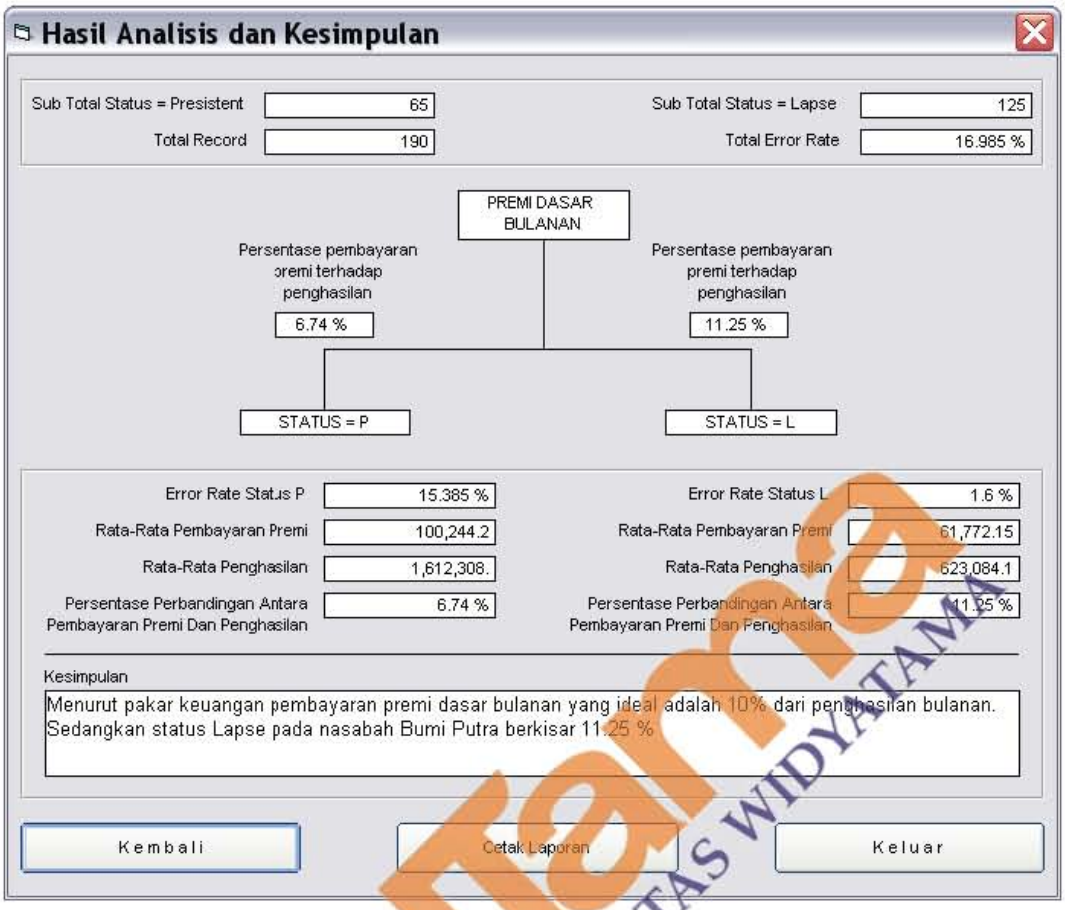
ISSN: 1907-5022

Gambar 1. Tampilan Menu Analisa Data Training



Gambar 2. Tampilan Menu Analisa Data Testing

ISSN: 1907-5022



Gambar 3. Tampilan Menu Analisa Data Uji

Kesimpulan

Berikut adalah kesimpulan yang dapat diambil dari data nasabah asuransi setelah dilakukan analisis menggunakan metode algoritma C4.5,

- Aplikasi dapat menyimpulkan bahwa rata-rata nasabah memiliki status L dikarenakan pembayaran premi yang melebihi 10% dari penghasilan.
- Dengan persentase atribut Premi_Dasar dan Penghasilan, maka dapat diketahui rata-rata status nasabah memiliki nilai P atau L.

Daftar Pustaka

[1] http://home.unpar.ac.id/~integral/Volume8/Integral8,No.2/C45Algorithm.PDF

[2] Kantardzic, Mehmed , Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003.

Larose, Daniel T., Discovering Knowledge in Data An Introduction to Data Mining, John Wiley & Sons,Inc., Hoboken, New Jersey., 2005