PENERAPAN DECISION TREE UNTUK MENGANALISIS KEMUNGKINAN PENGUNDURAN DIRI CALON MAHASISWA BARU

Andie, S.Kom., M.Kom
(andina777@gmail.com)

ABSTRAK

Siswa adalah salah satu aset penting dalam sebuah Universitas, terutama di perguruan tinggi swasta. Kecenderungan didapat mahasiswa baru untuk perguruan tinggi swasta di Indonesia telah menurun dan membuat proses pendaftaran mahasiswa baru lebih menantang dari pada sebelumnya. Perguruan tinggi kesulitan mendapatkan informasi mengenai jumlah mahasiswa baru yang akan registrasi dari daftar mahasiswa yang sudah dinyatakan lulus dalam seleksi penerimaan mahasiswa baru. Penelitian ini mencoba untuk memprediksi apakah seorang siswa calon akan cenderung melanjutkan studi mereka atau tidak, menggunakan metode Data mining Decision Tree yang disebut algoritma C4.5. Seperangkat aturan akan dibuat sebagai dasar untuk dibandingkan dengan data baru yang pada akhirnya akan menghasilkan prediksi apakah siswa akan melanjutkan studi (proses registrasi) atau tidak. Penelitian ini akan dilakukan di Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad Al-Banjary Banjarmasin menggunakan data 3 tahun yaitu tahun 2013 sampai 2015. Adapun hasil akhirnya, UNISKA akan dapat memprediksi (dari jumlah total terdaftar calon mahasiswa), kemungkinan setiap siswa yang akan melanjutkan proses sampai mereka belajar di UNISKA. Output dari Sistem ini akan dilaksanakan dalam Promosi Penerimaan Mahasiswa Baru tahun yang akan datang karena manajemen akan mengambil langkah lebih lanjut dalam mengantisipasi sasaran mahasiswa baru setiap tahun.

Kata Kunci : Data mining, Decision Tree

PENDAHULUAN

Tujuan diselenggarakanya seleksi penerimaan mahasiswa baru adalah untuk membentuk mahasiswa yang unggul dalam pengembangan ilmu pengetahuan, teknologi, dan seni, serta pelestarian dan pengkayaan budaya bangsa untuk mengembangkan manusia yang memiliki kecerdasan intelektual, emosional dan spiritual yang Islami sesuai dengan visi, misi dan tujuan dari Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad Al-Banjary Banjarmasin. Selain itu jurusan eksakta membutuhkan ketekunan dan keteraturan, sedangkan jurusan non eksakta membutuhkan kepekaan sosial. Minat akan mempengaruhi dalam semangat belajar dan menimbulkan dorongan rasa ingin tahu. Minat yang tepat harus disesuaikan dengan kemampuan dan kepribadian. Bukan karena pengaruh orang lain atau trend (Psikologi, 2010).

Penelitian ini mencoba untuk memprediksi apakah seorang siswa calon mahasiswa akan cenderung melanjutkan studi mereka atau tidak, menggunakan metode Data mining yang disebut algoritma C4.5 sebagai klasifikasi atribut pada dataset.

Seperangkat aturan akan dibuat sebagai dasar untuk dibandingkan dengan data baru yang pada akhirnya akan menghasilkan prediksi apakah siswa akan melanjutkan studi (proses registrasi) atau tidak.

Penelitian ini akan dilakukan di Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad Al-Banjary Banjarmasin menggunakan data pendaftaran mahasiswa baru selama 3 tahun, yaitu tahun 2013 sampai 2015. Menurut data mahasiswa baru yang diperoleh, pada tahun 2013 terdapat 16% calon mahasiswa yang memundurkan diri, pada tahun 2014 terdapat 15% calon mahasiswa yang memundurkan diri, sedangkan pada tahun 2015 terdapat 20% calon mahasiswa yang memundurkan diri.

Adapun hasil akhirnya, UNISKA akan dapat memprediksi (dari jumlah total terdaftar calon mahasiswa), berapa banyak kemungkinan setiap siswa yang akan melanjutkan proses sampai mereka belajar di UNISKA. Output dari Sistem ini akan dilaksanakan dalam Promosi Penerimaan Mahasiswa Baru tahun yang akan datang karena manajemen akan mengambil langkah lebih lanjut dalam mengantisipasi sasaran mahasiswa baru setiap tahun.

RUMUSAN MASALAH

Pihak perguruan tinggi kesulitan mengetahui jumlah mahasiswa yang akan registrasi dari sejumlah calon mahasiswa yang sudah dinyatakan lolos seleksi penerimaan mahasiswa baru. Dengan tidak diketahuinya mahasiswa yang akan registrasi, maka pihak-pihak perguruan tinggi tidak dapat mengetahui dengan pasti kapan jumlah target penerimaan mahasiswa baru tercapai. Hal ini dikarenakan jumlah mahasiswa yang telah dinyatakan lolos seleksi penerimaan mahasiswa baru, tidak menggambarkan jumlah mahasiswa yang akan melakukan registrasi. Apabila hal tersebut baru diketahui pada akhir proses penerimaan mahasiswa baru akan mengakibatkan target menjadi tidak tercapai. Hal ini juga dapat menjadi gangguan dalam hal finansial, karena biasanya pembiayaan perguruan tinggi swasta sebagian besar dari dana yang didapatkan dari mahasiswa.

TUJUAN PENELITIAN

Penelitian ini dilakukan dengan tujuan untuk membantu pihak perguruan tinggi swasta, khususnya dalam hal memprediksi jumlah kemungkinan mahasiswa yang akan melakukan registrasi ulang secepat mungkin. Hal tersebut dapat meningkatkan efektifitas target dan jumlah mahasiswa yang akan diterima serta untuk bahan kebijakan penerimaan siswa baru. Universitas akan dapat mengetahui lebih awal siswa mana yang akan mendaftar ulang dan yang memundurkan diri.

TARGET LUARAN DAN MANFAAT PENELITIAN

Target luaran yang akan dicapai dalam penelitian ini adalah Publikasi Ilmiah dalam jurnal lokal yang mempunyai ISSN atau jurnal nasional seperti jurnal Al Ulum Universitas Islam Kalimantan (UNISKA), Proseding pada seminar ilmiah baik yang berskala lokal, regional maupun nasional dan sebagai pengayaan bahan ajar.

METODE PENELITIAN

Jenis penelitian ini dilakukan secara ekperimen dalam melakukan pengujian terhadap data. Penelitian ini dilakukan dengan tujuan untuk membantu pihak perguruan tinggi swasta, khususnya dalam hal

memprediksi jumlah kemungkinan mahasiswa yang akan melakukan registrasi ulang secepat mungkin. Hal tersebut dapat meningkatkan efektifitas target dan jumlah mahasiswa yang akan diterima serta untuk bahan kebijakan penerimaan siswa baru. Universitas akan dapat mengetahui lebih awal siswa mana yang akan mendaftar ulang dan yang memundurkan diri. Jenis data yang digunakan dalam penelitian ini adalah data primer yang diperoleh dari panitia penerimaan siswa baru dari tahun 2013 sampai dengan 2015.

Penelitian menggunakan data mahasiswa baru di Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad Al-Banjary Banjarmasin selama 3 tahun yaitu antara tahun 2013 sampai tahun 2015. Data yang dibutuhkan dalam penelitian ini adalah:

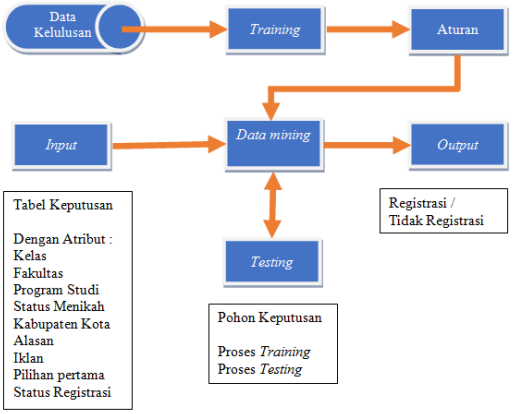
Data Primer, Data primer yang digunakan adalah data yang langsung diperoleh dari database penerimaan siswa baru. Yaitu data yang langsung diberikan oleh panitia penerimaan siswa baru.

Data Sekunder, Data sekunder yang digunakan adalah data yang sudah dimodifikasi dan dipisah pertahun serta perbaikan type data sehingga mudah untuk dibaca dan melakukan ekperimen.

Dari data primer yang ada akan dilakukan tahap penyesuain data diantara sebagai berikut :

1. Menghitung jumlah field awal
2. Menghapus field yang tidak diperlukan
3. Mengelompokkan type data
4. Pemilihan atribut

Metode penelitian yang dilakukan seperti yang tergambar pada Gambar 3.1. Proses awal adalah mengolah data kelulusan dan registrasi calon mahasiswa baru setahun terakhir. Data tersebut dijadikan data sampel untuk membentuk pohon keputusan yang akan membentuk aturan dengan menggunakan algoritma C4.5. Setelah aturan terbentuk, selanjutnya adalah memasukkan data calon mahasiswa baru yang sudah lolos seleksi untuk diprediksi dengan menggunakan aturan tadi.



Proses pembentukan pohon keputusan dalam pengelolaan data digunakan untuk menurunkan atribut-atribut pada data menjadi pohon keputusan sehingga menghasilkan aturan-aturan. Misalnya terdapat data sampel mahasiswa dari sebuah universitas, kemudian dari data tersebut dipilih atribut-atribut yang bisa mewakili atribut lain yang ada pada data untuk dicari pola atau informasi yang ada di dalamnya. Karena keterbatasan sampel data yang digunakan, maka terpilihlah 9 atribut yang akan digunakan untuk pembentukan pohon keputusan.

Penurunan atribut menjadi pohon keputusan menggunakan algoritma C4.5 dilakukan dengan konsep Entropy dan Information gain. Konsep ini digunakan untuk menentukan node induk dan node daun dalam pohon keputusan. Sebuah obyek yang diklasifikasikan dalam pohon harus dites nilai entropinya. Entropy adalah ukuran dari teori informasi yang dapat mengetahui karakteristik dari impury, dan homogenity dari kumpulan data. Dari nilai Entropy tersebut kemudian dihitung nilai Information gain (IG) masing-masing atribut.

Pemillihan atribut pada C4.5 dilakukan dengan properti statistik, yang disebut dengan Information gain. Gain mengukur seberapa baik suatu atribut memisahkan training example ke dalam kelas target. Atribut dengan informasi tertinggi akan dipilih. Dengan tujuan untuk mendefinisikan gain, pertama-tama digunakanlah ide dari teori informasi yang disebut Entropy. Entropy mengukur jumlah dari informasi yang ada pada atribut. Setelah mendapat nilai Entropy untuk suatu kumpulan data, maka kita dapat mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Ukuran efektifitas ini disebut Information gain.

Pada penelitian dengan seluruh data tidak mungkin dilakukan secara manual, maka peneliti melakukan eksperimen menggunakan aplikasi Rapid Miner versi 5. Data yang dibunakan sebanyak 10663 data. Adapun tahapan eksperimen adalah sebagai berikut :

1. Input data melalui Read Excel
2. Melakukan Import Configuration Wizart
3. Memilih data mahasiswa baru yang berexstensi excel (.xlsx)
4. Memilih atribut dan label
5. Setelah data berhasil diimport maka proses eksperimen data mahasiswa akan dapat dilakukan.
6. Membuat New Building Blok dan memilih Nominal X Validation
7. Memasukkan Decision Tree pada proses Training, Apply Model dan Performance pada proses Testing
8. Menghubungkan jalur benang data sesuai yang dikehendaki.

Selain proses diatas peneliti juga melakukan pengujian data menggunakan proses dibawah ini :

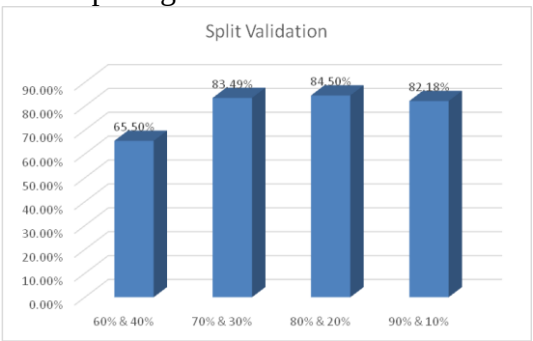
1. Proses Training dan Testing dengan cara membagi data menjadi 2 dengan 4 ketentuan yaitu :
   a. Membagi data 60% dan 40%
   b. Membagi data 70% dan 30%
   c. Membagi data 80% dan 20%
   d. Membagi data 90% dan 10%
2. Melakukan sampling type data yaitu :
   a. Linier Sampling
   b. Shuffled Sampling
   c. Stratified Sampling

Selanjutnya dilakukan perbandingan antara algoritma C4.5 dengan algoritma ID3. Diharapkan penelitian ini akan dapat mengetahui perhitungan algoritma mana yang lebih bagus untuk digunakan untuk menghitung akurasi tertinggi pada data calon mahasiswa baru.

ANALISA HASIL DAN PEMBAHASAN

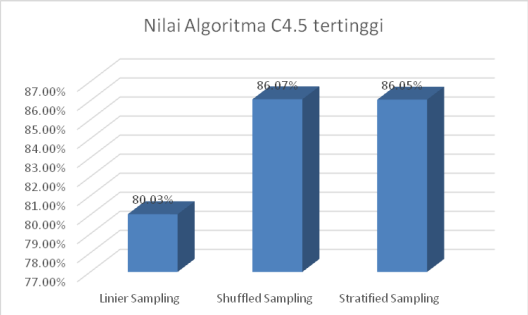
1. Split Validation

Proses training dan Testing peneliti lakukan dengan cara membagi data menjadi 4 karegori yaitu 60% - 40%, 70% - 30%, 80% - 20% dan 90% - 10%. Hasil dari penelitian tersebut dapat dilihat pada grafik dibawah ini:

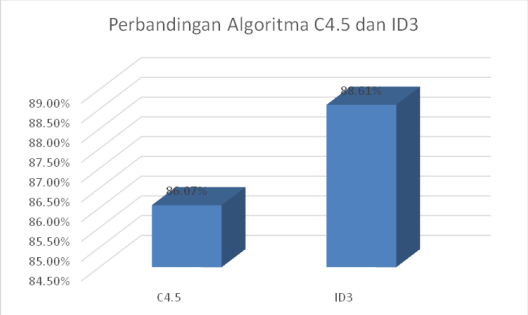


2. Metode Pengujian Sampel

Pengujian sampel pada penelitian ini ada 3 yaitu, Linier Sampling, Shuffled Sampling dan Stratified Sampling. Hasil pengujian sampel kemudian dibandingkan lagi menggunakan X-validation dengan Number of Vaidation dari 2 sampai 10. Hasil dari penelitian tersebut dapat dilihat pada grafik dibawah ini:

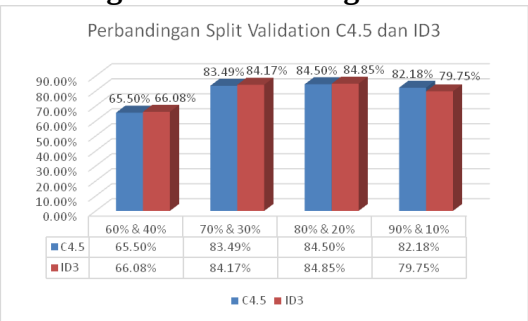


3. Perbandingan Algoritma C4.5 dan ID3



Dari penelitian diatas dapat dilihat bahwa algoritma ID3 lebih unggul dari algoritma C4.5 dengan akurasi 88.61% banding 86.07%.

4. Perbandingan Split Validation Algoritma C4.5 dengan ID3



Dari hasil perbandingan diatas dapat dilihat bahwa algoritma *ID3* tetap masih unggul dibandingkan dengan algoritma *C4.5* dengan hasil perbandingan 84.85% banding 84.50%.

PENUTUP

1. Kesimpulan

Penelitian ini dilakukan pada Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad Al-Banjary Banjarmasin dengan sampel data yang digunakan sebanyak 10663 data calon mahasiswa baru tahun 2013, 2014 dan 2015. Penelitian dilakukan dengan metode Data mining yang disebut Decision Tree (C4.5).

Penelitian pertama dilakukan dengan perhitungan manual serta menggunakan sampel data sebanyak 35 data dan menghasilkan akurasi sebanyak 94.29%. Penelitian kedua dilakukan dengan menggunakan aplikasi Rapid Miner serta menggunakan sampel data keseluruhan yaitu sebanyak 10663 data dan menghasilkan akurasi akhir tertinggi sebanyak 86.07%. Penelitian ketiga dilakukan dengan cara membandingkan antara algoritma C4.5 dengan ID3 dengan hasil perhitungan algoritma ID3 lebih unggul dibandingkan dengan algoritma C4.5. Dan terakhir penelitian ketiga dilakukan dengan pemisahan data Training dan Testing dengan menghasilkan akurasi tertinggi 84.50%.

2. Saran

Kekurangan pada penelitian ini adalah karena keterbatasan atribut yang dapat digunakan, diharapkan bagi yang dapat meneruskan untuk sekiranya meningkatkan hasil akurasi dengan atribut yang lebih banyak lagi.

DAFTAR PUSTAKA

Al-Radaideh, Q., Al-Shawakfa, E. & Al-Najjar, M., 2006. *Mining Student Data Using Decision Trees.* Arab Saudi: International Arab Conference on Information Technology (ACID).

Craw, S., 2014. *Case Based Reasoning.* [Online] Available at: www.comp.rgu.ac.uk/staff/smc/teaching/cm3016/Lecture-3-cbr-indexing.ppt

Defiyanti, S., Pardede & Crispina, D., 2014. *Perbandingan Kinerja Algoritma ID3 dan C4.5 dalam Klasifikasi Spam-Mail.* s.l.:Jurnal Jurusan Sistem Informasi Universitas Gunadarma.

Fayyad, U., 1996. *Advanced in Knowledge Discovery and Data Mining.* s.l.:MIT Press.

Han, J. & Kamber, M., 2001. *Data Mining : Concept and Techniques.* Sun Fransisco: Morgan Kaufmann Publishers.

Junanto, A., 2010. Perbandingan Performansi Algoritma Nearest Neighbor dan SLIQ untuk Prediksi Kinerja Akademik Mahasiswa Baru. *Jurnal Teknologi Informasi DINAMIK,* Volume XV, pp. 157-169.

Kaller, D. & Pierrakeas, C., 2006. *Analyzing Student Performance in Distance Learning With Genetic Algorithms and Decsion Trees.* s.l.:Hellenic Open University.

Kusrini & Hartati, S., 2007. Penggunaan Pohon Keputusan Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Di Stmik Amikom Yogyakarta. *Seminar Nasional Teknologi 2007,* Volume 7, p. 5.

Larose & Daniel, T., 2005. *Discovering Knowledge in Data : An Introduction to Data Mining.* s.l.:John Willey & Sons. Inc.

Pramudyo, A. S., 2008. *Case Based Reasoning untuk Klasifikasi Mahasiswa Baru Berdasarkan Prediksi Indeks Prestasi pada Semester I (Studi Kasus Program Studi Teknik Informatika Universitas Bina Darma Palembang).* Yogyakarta: Tesis, Jurusan Ilmu Komputer FMIPA UGM.

Prasetia, K., 2008. *Pengertian Basis Data.* [Online] Available at: https://krida85.wordpress.com/2008/04/16/pengertian-basis-data/

Psikologi, U. P., 2010. *Tes Psikologi Pemilihan Jurusan di Perguruan Tinggi.* s.l.:s.n.

Pudjiantoro, T. H., Renaldi, F. & Teugunadi, A., 2011. Penerapan Data Mining untuuk Menganalisa Kemungkinan Pengunduran Diri Calon Mahasiswa Baru. *Konferensi Nasional Sistem dan Informatika 2011,* Volume 9, p. 59.

Santoso, B., 2007. Teknik Pemanfaatan Data untuk. In: *Data Mining.* Yogyakarta: Graha Ilmu.

Sembiring, A. P. U. & Ginting, M., 2013. Analisis Faktor-Faktor Yang Mempengaruhi Pengunduran Diri Mahasiswa Dengan Aplikasi Data Mining Add-Ins. *JSM STMIK Mikroskil,* Volume 14, p. 146.

Sugianti, D., 2012. Algoritma Bayesian Classification Untuk Memprediksi Heregristrasi Mahasiswa Baru Di STMIK Widya Pratama. *Jurnal Ilmiah ICTech,* Volume X, p. 5.

Wahyudin, 2011. *Metode Iterative Dichotomizer 3 (ID3) Untuk Penerimaan Mahasiswa Baru.* [Online] Available at: http://file.upi.edu/Direktori/FPMIPA/PRODI._ILMU_KOMPUTER/WAHYUDIN/metode_ID3_untuk_mhs_baru.pdf [Accessed 20 Januari 2015].

Wibowo, M. N., 2012. *Implementasi dan Demo Pohon Keputusan ID3 dan C4.5.* [Online] Available at: http://nugikkool.blogspot.com/2012/08/pohon-keputusan-id3-dan-c45-menggunakan.html [Accessed 15 Januari 2015].

Yan, H., Ma, R. & Tong, X., 2005. *SLIQ in data mining and application in*

the generation unit's bidding decision system of electricity market. s.l.:International Power Engineering Conference (IPEC2005).