ISSN: 1907-7912 EISSN: 2622-8092

PENERAPAN ALGORITMA C4.5 DALAM PEMILIHAN KONSENTRASI PROGRAM STUDI (Studi Kasus di STMIK Widya Pratama Pekalongan)

Tri Agus Setiawan, Agus Ilyas
STMIK Widya Pekalonagn Pekalongan
Jl. Patriot Kota Pekalongan
tri.triagus.setiawan45@gmail.com, agusilyas@gmail.com

Abstrak

Pemilihan konsentrasi program studi oleh mahasiswa sangatlah menentukan kompetensi mahasiswa setelah lulus, banyak faktor yang menjadi pertimbangan mahasiswa baik faktor intern maupun ekstern dari mahasiswa itu sendiri. Oleh karena itu dalam penelitian ini akan diklasifikasikan faktor apa yang menentukan mahasiswa dalam menentukan pemilihan konsentrasi pada tiap-tiap program studi yang ada di STMIK Widya Pratama yang meliupti variabel Kurikulum, Proram Studi, Citra Perguruan Tinggi, Kinerja (performance) Lulusan dan Peluang Kerja, Biaya. Dalam penelitian yang dilakukan menggunakan algoritma C4.5 agar dapat membantu dalam pengklasifikasian variable-variabel yang mempengaruhi pemilihan konsentrasi program studi. Algortima C4.5 merupakan algoritma yang cukup efektif dalam membantu membentuk sebuah pohon keputusan, pohon keputusan tersebut kemudian akan menghasilkan sebuah pengetahuan baru. Berdasarkan dari hasil pengujian yang dilakukan terhadap pohon keputusan diperoleh hasil bahwa faktor yang menentukan mahasiswa dalam memilih konsentrasi program studi sebesar 91% berdasarkan variable Kurikulum.

Kata kunci *: Konsentrasi Program Studi, Algoritma C4.5, Pohon Keputusan*

1. Pendahuluan

1.1. Latar belakang

Pemilihan konsentrasi program studi oleh mahasiswa pada awal perkuliahan sangat menentukan jenis bidang ilmu apa yang diminati karena akan disesuaikan dengan kemampuan dan kompetensi yang dimiliki mahasiswa setelah lulus. Bidang peminatan merupakan salah satu bagian dari kurikulum berbasis kopetensi yang berisi kumpulan dari beberapa matakuliah yang harus diambil mahasiswa menuju proses penyelesaian Tugas Akhir/Skripsi. STMIK Widya Pratama Pekalongan pada saat ini memiliki program studi Teknik Informatika konsentrasi (*graphic and multimedia softare, mobile aplication*), Sistem Informasi (*e-buiness, business intelligent system)* (S1) dan Manajemen Informatika (*programming and database administration*), Komputerisasi Akuntansi (D3).

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis yang menemukan keteraturan, pola dan hubungan dalam set data berukuran besar (Witten, 2011). Dalam perkembangan data mining ada 10 algoritma teratas yang paling berpengaruh yang dipilih oleh peneliti dalam komunitas data mining, dimana 6 (enam) diantaranya adalah algoritma classification (klasifikasi) yaitu C4.5, Support Vector Machines (SVM), AdaBoost, *K*-Nearest Neighbor (*kNN*), Naïve Bayes dan CART (Wu, Kumar, 2009). Salah satu algoritma yang pada saat ini dilakukan penelitaian adalah algoritma klasifikasi C4.5 (Lusinia, 2014).

Kelebihan dari algoritma C4.5 adalah kemudahan dalam pengambilan keputusan bagi pengembangan program studi berdasarkan klasifikasi kompetensi mahasiswa yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik. Adpun kekurangannya adalah hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana mahasiswa dalam memilih kompetensi program studi yang diinginkan.

Berdasarkan latar belakang masalah diatas maka diusulkan metode algoritma C4.5 untuk mengetahui klasifikasi sebaran pemilihan kompetensi program studi berdasarkan kriteria ekonomi, orang tua, jenis pekerjaan yang diinginkan, kebutuhan pangsa pasar dan yang lainnya Hal ini bertujuan untuk membantu pengambilan keputusan tentang pengembangan kompetensi program studi baik pengembangan kurikulum, pengemban dosen, sarana dan

prasarana pendukung lainynnya sesuai dengan perkembangan teknologi informasi dan kebutuhan pasar tenaga kerja

1.2. Landasan Teori

1.2.1. Pohon Keputusan (Decision *Tree*)

Decision Tree adalah flow-chart seperti struktur tree, dimana tiap internal node menunjukkan sebuah test pada sebuah atribut, tiap cabang menunjukkan hasil dari test dan leaf node menunjukkan class-class atau class distribution (Sunjana, 2010).

1.2.2. Algoritma C4.5

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 kebenarannya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data (Sunjana, 2010). Ada tiga prinsip kerja algoritma C4.5 pada tahap belajar dari data, yaitu sebgai berikut:

1. Pembuatan Pohon Keputusan
2. Pemangkasan Pohon Keputusan dan Evaluasi (Opsional)
3. Pembuatan Aturan-Aturan dari Pohon Keputusan (Opsional)

Algoritma C4.5 dapat menangani data numerik dan diskret. Algoritma C4.5 menggunakan rasio perolehan (*gain ratio*). Sebelum menghitung rasio perolehan, perlu dilakukan perhitungan nilai informasi dalam satuan bits dari suatu kumpulan objek, yaitu dengan menggunakan konsep entropi.

a. Langkah Konstruksi Pohon Keputusan dengan Algoritma C4.5
   1) Pohon dimulai dengan sebuah simpul yang mereperesentasikan sampel data pelatihan yaitu dengan membuat simpul akar.
   2) Jika semua sampel berada dalam kelas yang sama, maka simpul ini menjadi daun dan dilabeli menjadi kelas. Jika tidak, *gain ratio* akan digunakan untuk memilih atribut split, yaitu atribut yang terbaik dalam memisahkan data sampel menjadi kelas-kelas individu.
   3) Cabang akan dibuat untuk setiap nilai pada atribut dan data sampel akan dipartisi lagi.
   4) Algoritma ini menggunakan proses rekursif untuk membentuk pohon keputusan pada setiap data partisi. Jika sebuah atribut sduah digunakan disebuah simpul, maka atribut ini tidak akan digunakan lagi di simpul anak-anaknya.
   5) Proses ini berhenti jika dicapai kondisi seperti berikut :
      - Semua sampel pada simpul berada di dalam satu kelas
      - Tidak ada atribut lainnya yang dapat digunakan untuk mempartisi sampel lebih lanjut. Dalam hal ini akan diterapkan suara terbanyak. Ini berarti mengubah sebuah simpul menjadi daun dan melabelinya dnegan kelas pada suara terbanyak.

b. Konsep Entropi

Entropi (S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S. Entropi dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai entropi maka akan semakin entropy digunakan dalam mengekstrak suatu kelas. Entropi digunakan untuk mengukur ketidakaslian S. Besarnya Entropi pada ruang sampel S didefinisikan dengan:

$$Entropy(X) = \sum_{j=1}^{k} p_j * log_2 \frac{1}{p_j} = -\sum_{j=1}^{k} p_j * log_2 p_j$$

dimana,
X: Himpunan Kasus
k: jumlah partisi X
p_j: Proporsi X_j terhadap X

Entropi split yang membagi *X* dengan *n* record menjadi himpunan-himpunan X_1 dengan n_1 baris dan X_2 dengan n_2 baris adalah:

$$E(X_1, X_2) = \frac{n1}{n} E(X_1) + \frac{n2}{n} E(X_2)$$

Besar nilai *Entropy(X)* menunjukkan bahwa X adalah atribut yang lebih acak. Di sisi lain, atribut yang lebih kecil dari nilai *Entropy(X)* menyiratkan atribut ini sedikit lebih acak yang signifikan untuk data mining. Nilai entropi mencapai nilai minimum 0, ketika semua p_j lain = 0 atau berada pada kelas yang sama. Nilainya mencapai maksimum $\log_2$ k, ketika semua nilai p_j adalah sama dengan 1/k.

c. Gain Ratio

Pada kontruksi pohon C4.5, di setiap simpul pohon, atribut dengan nilai *gain ratio* tertinggi dipilih sebagai atribut split untuk simpul. Rumus dari gain ratio adalah sebagai berikut :

$$gain\ ratio(a) = \frac{gain(a)}{split(a)}$$

Dimana *gain(a)* adalah *information gain* dari atribut *a* untuk himpunan sampel *X* dan *split info(a)* menyatakan entropi atau informasi potensial yang didapat pada pembagian *X* menjadi *n sub* himpunan berdasarkan telaahan pada atribut *a*. Sedangkan *gain(a)* didefinisikan sebagai berikut :

$$gain(a) = info(X) - info_a(X)$$

Untuk rumus *split info(a)* adalah sebagai berikut :

$$split\ info(a) = -\sum_{j=1}^{k} \frac{|X_1|}{|X|} * \log_2\left(\frac{|X_1|}{|X|}\right)$$

dimana X_i menyatakan sub himpunan ke-*I* pada sampel *X*.

Dengan kata lain rumus untuk menghitung nilai gain ratio untuk dipilih sebagai atribut dari simpul yang ada sebagai berikut ini :

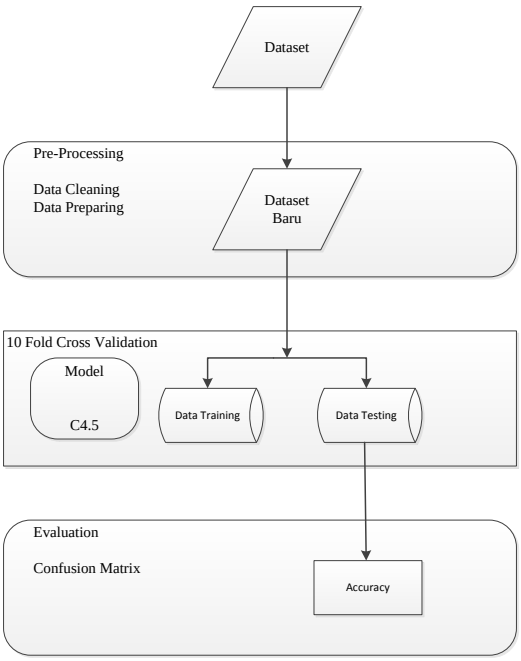
$$Gain\ ratio(a) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i)$$

Alasan penggunaan *gain ratio(a)* pada C4.5 (bukan *gain(a)*) sebagai kriteria pada pemilihan atribut adalah gain ternyata bias terhadap atribut yagn memiliki banyak nilai unik.

2. Metodologi Penelitian

Pada penelitian yang dilakukan menggunakan algoritma klasifikasi C4.5 untuk memetakan peminatan mahasiswa dalam memilih kompetensi program studi yang diinginkan dengan cara dilakukan penilaian atau pembobotan pada setiap kriteria yang ada.

Pada Gambar 3.3 menunjukan proses algoritma C4.5. Pertama dilakukan pengumpulan dataset, kemudian diproses menggunakan algo, kemudian untuk tahap terakhir dilakukan proses validasi.



Gambar 1. Metode yang Diusulkan

2.1. Dataset

Dalam penelitian ini data yang dihasilkan berdasarkan isian kuesioner yang dibagikan kepada seluruh mahasiswa semester 1 program studi D-III Manajemen Informatika, Komputerisasi Akuntansi dan S-1 Teknik Informatikan dan Sistem InformasiTahun Akademik 2018-2019 sebanyak 306 data

2.2. Pengujian

Dalam proses pengujian yang dilakukan dilakukan beberapa tahapan antara lain:

1. Data Selection

Variable yang dipakai dalam pemilihan konsentrasi program studi adalah Kurikulum, Proram Studi, Citra Perguruan Tinggi, Kinerja (performance) Lulusan dan Peluang Kerja, Biaya. Data penelitian yang dipakai seperti terlihat pada Tabel 1.

Tabel.1 Data Penelitian

NO	KONSENTRASI	KURIKULUM	PROGRAM STUDI	CITRA PERGURUAN TINGGI	KINERJA LULUSAN DAN PELUANG KERJA	BIAYA	MINAT
1	KA1	Ya	Ya	Ya	Ya	Tidak	Ya
2	KA2	Ya	Ya	Tidak	Ya	Tidak	Ya
3	KA3	Ya	Ya	Tidak	Tidak	Tidak	Ya
4	KA4	Tidak	Tidak	Tidak	Ya	Tidak	Tidak
5	KA5	Ya	Tidak	Ya	Tidak	Ya	Ya
6	KA7	Ya	Ya	Ya	Ya	Ya	Ya
7	PDA1	Ya	Tidak	Tidak	Tidak	Tidak	Tidak
8	PDA2	Ya	Ya	Ya	Tidak	Tidak	Ya
9	PDA3	Ya	Ya	Ya	Ya	Ya	Ya
10	PDA5	Ya	Ya	Ya	Ya	Tidak	Ya
11	PDA10	Ya	Ya	Tidak	Ya	Ya	Ya
12	PDA21	Ya	Tidak	Tidak	Ya	Tidak	Tidak
13	SIEB2	Ya	Ya	Ya	Ya	Ya	Ya
14	SIEB3	Ya	Ya	Ya	Tidak	Tidak	Ya
15	SIEB8	Ya	Ya	Ya	Tidak	Ya	Ya
16	SIEB23	Ya	Tidak	Ya	Ya	Tidak	Ya
17	SIEB24	Ya	Ya	Tidak	Ya	Tidak	Ya
18	SIEB27	Tidak	Ya	Ya	Ya	Tidak	Ya
19	SIEB29	Ya	Tidak	Ya	Ya	Tidak	Ya
20	SIEB30	Tidak	Ya	Tidak	Ya	Tidak	Tidak
21	SIBIS7	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
22	SIBIS9	Tidak	Tidak	Ya	Ya	Tidak	Tidak
23	SIBIS24	Ya	Ya	Tidak	Ya	Tidak	Ya
24	SIBIS26	Ya	Ya	Ya	Ya	Tidak	Ya
25	SIBIS34	Ya	Tidak	Tidak	Ya	Tidak	Tidak
26	SIBIS35	Ya	Ya	Tidak	Ya	Ya	Ya
27	SIBIS54	Tidak	Ya	Ya	Ya	Tidak	Ya
28	SIBIS78	Ya	Ya	Tidak	Tidak	Tidak	Tidak
29	TIMA1	Ya	Tidak	Tidak	Tidak	Tidak	Tidak
30	TIMA3	Ya	Ya	Ya	Ya	Ya	Ya
31	TIMA6	Ya	Tidak	Ya	Tidak	Tidak	Tidak
32	TIMA7	Ya	Ya	Tidak	Ya	Tidak	Ya
33	TIMA9	Ya	Ya	Tidak	Ya	Ya	Ya
34	TIMA11	Tidak	Tidak	Tidak	Ya	Tidak	Tidak
35	TIMA18	Ya	Tidak	Tidak	Ya	Tidak	Tidak
36	TIMA22	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
37	TIMA26	Tidak	Ya	Tidak	Tidak	Tidak	Tidak
38	TIMA27	Ya	Ya	Ya	Ya	Tidak	Ya
39	TIMA32	Ya	Tidak	Ya	Ya	Tidak	Ya
40	TIMA35	Ya	Ya	Ya	Tidak	Tidak	Ya
41	TIMA36	Ya	Tidak	Ya	Ya	Ya	Ya
42	TIMA37	Tidak	Ya	Ya	Ya	Ya	Ya
43	TIMA40	Ya	Tidak	Ya	Tidak	Ya	Ya
44	TIGMS1	Ya	Ya	Tidak	Ya	Ya	Ya
45	TIGMS4	Ya	Ya	Tidak	Ya	Tidak	Ya
46	TIGMS6	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
47	TIGMS7	Ya	Tidak	Tidak	Tidak	Tidak	Tidak
48	TIGMS9	Ya	Ya	Ya	Ya	Tidak	Ya
49	TIGMS10	Ya	Ya	Tidak	Ya	Tidak	Ya
50	TIGMS25	Ya	Tidak	Tidak	Ya	Tidak	Tidak
51	TIGMS32	Ya	Ya	Tidak	Tidak	Tidak	Tidak
52	TIGMS39	Ya	Tidak	Tidak	Ya	Tidak	Tidak
53	TIGMS40	Tidak	Ya	Tidak	Ya	Tidak	Tidak
54	TIGMS47	Ya	Tidak	Tidak	Tidak	Ya	Tidak

2. Transformation

Proses transformasi yang dilakukan adalah mengklasifikasikan Minat pemilihan konsentrasi menjadi 2 label yaitu “Ya” untuk Minat >= 3 variabel dan “Tidak” untuk Minat < 3 variabel. Hasil transformasi dapat dilihat pada Tabel 2.

Tabel.2 Data Transformasi

Ketegori	**Keterangan**
F1	Kurikulum
F2	Program Studi
F3	Citra Perguruan Tinggi
F4	Kinerja Lulusan dan Peluang Kerja
F5	Biaya
Ya	1
No	0
Minat Ya	>= 3
Minat Tidak	< 3

3. Hasil dan Pembahasan

3.1. Penerapan Algoritma C4.5

Berdasarkan data hasil transformasi kemudian dianalisa untuk dapat menghasilkan sebuah pohon keputusan dengan algoritma C4.5, secara umum untuk membangun pohon keputusan adalah sebagai berikut:

1. Perhitungan *Entropy* dan *Gain*
2. Pemilihan Gain tertinggi sebagai akar (Node)
3. Ulangi proses perhitungan Entropy dan Gain untuk mencari cabang sampai semua kasus pada cabang memiliki kelas yang sama yaitu pada saat semua variabel telah menjadi bagian dari pohon keputusan atau masing – masing variabel telah memiliki daun atau keputusan.
4. Membuat Rule berdasarkan pohon keputusan.

Untuk dapat memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus sebagai berikut:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si)$$

Di mana :

1. *S* : Himpunan Kasus
2. *A* : *Atribut*
3. *n* : *Jumlah Partisi Atribut A*
4. *|Si|* : *Jumlah Kasus pada Partisi ke-i*
5. *|S|* : *Jumlah* Kasus dalam S

Perhitungan nilai *entropy* dapat dilihat pada persamaan berikut ini:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

Di mana :

1. *S* : Himpunan Kasus
2. *A* : Atribut
3. *n* : Jumlah Partisi *S*
4. p_i : Proporsi dari S_i terhadap *S*

3.2. Pengujian

1. Entropy

Entropy digunakan untuk mengukur ketidakaslian S. Besarnya Entropi pada ruang sampel S didefinisikan dengan:

$$Entropy(X) = \sum_{j=1}^{k} p_j * log_2 \frac{1}{p_j} = -\sum_{j=1}^{k} p_j * log_2 p_j$$

= 0,863860036

2. Gain Ratio

Pada kontruksi pohon C4.5, di setiap simpul pohon, atribut dengan nilai *gain ratio* tertinggi dipilih sebagai atribut split untuk simpul.

$$gain(a) = info(X) - info_a(X)$$

= 0,133683366

$$split\ info(a) = -\sum_{j=1}^{k} \frac{|X_1|}{|X|} * \log_2\left(\frac{|X_1|}{|X|}\right)$$

= 0,133683366

$$gain\ ratio(a) = \frac{gain(a)}{split(a)}$$

= 0,05757

3. Pohon Keputusan

Dari hasil pengujian yang dilakukan terhadap mahasiswa dalam memilih konsentrasi program studi terhadap mahasiswa semester 1 tahun akademik 2018-2019 seperti pada Tabel 3 diperoleh hasil bahwa 91% mahasiswa memilih konsentrasi program studi berdasarkan variable Kurikulum.

Tabel 3 Tabel Frekuensi dan Prosentase Mahasiswa dalam Memilih Konsentrasi Program Studi

1. Kurikulum	277	29	306	91%	9%	100%
2. Program Studi	266	40	306	87%	13%	100%
3. Citra Perguruan Tinggi	223	83	306	73%	27%	100%
4. Kinerja (performance) Lulusan dan Peluang Kerja	271	35	306	89%	11%	100%
5. Biaya	250	56	306	82%	18%	100%

4. Kesimpulan

Dari hasil penelitian yang dilakukan dapat diambil kesimpulan berdasarkan hasil perhitungan dengan algoritma C4.5 diperoleh hasil mahasiswa memilih konsentrasi program studi berdasarkan kurikulum yaitu sebesar 91%

5. Saran dan Penelitian Selanjutnya

Untuk penelitian selanjutnya perlu juga penambahan variabel hasil nilai ujian masuk, sehingga pemilihan konsentrasi program studi dapat juga dilihat berdasarkan peringkat nilai.

Daftar Pustaka:

Betha, Sidik, 2005, MySQL untuk Pengguna Administrator dan Pengembangan Aplikasi Web, Informatika, Bandung

Lusinia, Shary Armonitha, Algoritma C4.5 Dalam Menganalisa Kelayakan Kredit (Studi Kasus di Koperasi Pegawai Republik Indonesia (KP-RI) Lengayang Pesisir Selatan, Painan, Sumatera Barat), Jurnal KomTekInfo Fakultas Ilmu Komputer, Volume 1, No. 2, Desember 2014, SN : 2356-0010.

M. A. Witten, I. H., Frank, E., & Hall, *Witten - Data mining 3rd - 2011*, Third Edit. USA: Morgan Kaufmann Publishers, 2011, p. 665.

Pressman, Roger S, Rekayasa Perangkat lunak, edisi 7, 2012 Yogyakarta: Andi Offset.

R. L. Maimon Oded, *Data Mining And Knowledge Discovery Handbook*, Second Edi. Israel: Springer, 2010, p. 1306.

Sunjana, 2010. Seminar Nasional Aplikasi Teknologi Informasi 2010. Snati 2010. Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree, 24-29.

V. Wu, Xindong & Kumar, *The Top Ten Algorithm in Data Mining*. Boca Raton: Taylor & Francis Group, LLC, 2009