# PENERAPAN ALGORITMA C4.5 PADA PROGRAM KLASIFIKASI MAHASISWA *DROPOUT*

Anik Andriani
AMIK BSI Jakarta

ABSTRAK. Prestasi akademik mahasiswa dievaluasi setiap akhir semester untuk mengetahui hasil belajar yang telah dicapai. Apabila mahasiswa tidak dapat memenuhi kriteria akademik tertentu untuk dinyatakan layak melanjutkan studi maka mahasiswa tersebut dinyatakan putus kuliah atau *dropout* (DO). Salah satu faktor penyebab banyaknya jumlah mahasiswa DO karena kurangnya kebijakan dan tindakan dari instansi pendidikan untuk menjaga mahasiswanya dari DO. Tujuan Penelitian ini adalah membuat klasifikasi mahasiswa DO dan aktif dengan algoritma C4.5 sebagai acuan dalam membuat kebijakan dan tindakan untuk mengurangi jumlah mahasiswa DO. Hasil klasifikasi dari algoritma C4.5 dievaluasi dan divalidasi dengan *confusion matrix* dan kurva *Receiver Operating Characteristic* (ROC) untuk mengetahui tingkat akurasi Algoritma C4.5 dalam membuat klasifikasi mahasiswa potensi DO. Hasil evaluasi dan validasi diperoleh tingkat akurasi sebesar 97,75%. *Rule* yang diperoleh dari klasifikasi dengan Algoritma C4.5 jika diterapkan dalam data baru diperoleh hasil validasi dengan tingkat akurasi 90,0%.

***Kata Kunci:*** *Dropout, Klasifikasi, Algoritma C4.5*

## 1. PENDAHULUAN

Mahasiswa merupakan salah satu substansi yang perlu diperhatikan dalam kaitannya dengan dunia pendidikan, karena mahasiswa merupakan penerjemah terhadap dinamika ilmu pengetahuan, dan melaksanakan tugas yaitu mendalami ilmu pengetahuan tersebut, Harahap[1].Sebagai sebuah subjek yang berpotensi dan sekaligus objek dalam aktifitas dan kreatifitasnya, mahasiswa diharapkan mampu mengembangkan kualitas dirinya, Baharudin dan Makin[2].Kualitas tersebut dapat dilihat dari prestasi akademik yang diraihnya yang merupakan bukti usaha yang diperoleh mahasiswa, Sobur[3].

Prestasi akademik mahasiswa dievaluasi setiap akhir semester untuk mengetahui hasil belajar yang telah dicapai.Apabila mahasiswa tidak dapat memenuhi kriteria akademik tertentu untuk dinyatakan layak melanjutkan studi maka mahasiswa tersebut dinyatakan putus kuliah atau *dropout* (DO).Tingginya jumlah mahasiswa *dropout* pada perguruan tinggidapat diminimalisir dengan kebijakan dari perguruan tinggi untuk mengarahkan dan mencegah mahasiswa dari *dropout* seperti yang diungkapkan oleh Deker, Pechenizkiy, dan Vleeshouwer[4] bahwa mendeteksi mahasiswa berisiko pada tahap awal pendidikan sangat penting dilakukan untuk menjaga mahasiswa dari *dropout*.Hal ini memungkinkan departemen penyelenggara pendidikan untuk memberikan pengarahan kepada mahasiswa yang membutuhkan.Belum adanya sebuah alat untuk mendeteksi mahasiswa yang berpotensi *dropout* secara otomatis dapat mempersulit perguruan tinggi dalam membuat kebijakan yang tepat.

Alat untuk mendeteksi mahasiswa yang berpotensi *dropout*dapat dibangun dengan menerapkan hasil klasifikasi mahasiswa yang berpotensi aktif maupun DO.Klasifikasi

(taksonomi) adalah proses menempatkan suatu objek atau konsep ke dalam satu set kategori berdasarkan objek atau konsep yang bersangkutan, Gorunescu[5]. Klasifikasi juga dapat diartikan sebagai sebuah proses menemukan suatu model atau fungsi yang menggambarkan dan membedakan klas data atau konsep dengan tujuan dapat menggunakan model untuk membuat prediksi kelas objek dimana kelas labelnya tidak diketahui, Han dan Kamber[6]. Algoritma yang dapat digunakan untuk klasifikasi antara lain C4.5.

Identifikasi permasalahan dalam penelitian ini dapat dirumuskan dengan *research question* yaitu Apakah algoritma C4.5 dapat mengklasifikasi mahasiswa yang berpotensi *dropout*dengan tingkat akurasi yang tinggi?

Penelitian ini bertujuan untuk membuat klasifikasi mahasiswa yang berpotensi *dropout* dengan algoritma C4.5 dan menerapkan hasil klasifikasi dalam sebuah aplikasi untuk prediksi mahasiswa*dropout*.

## 2. TINJAUAN PUSTAKA

Berdasarkan sudut pandang operasional, data mining adalah proses terpadu dari analisis data yang terdiri dari serangkaian kegiatan yang berjalan berdasarkan definisi tujuan yang akan dianalisis, dengan analisis datanya sampai interpretasi dan evaluasi hasil, Guidici dan Figini [7].

Data mining didefinisikan sebagai proses menemukan pola dalam data. Proses ini harus otomatis atau biasanya secara semi-otomatis.Pola yang dihasilkan harus berarti bahwa pola tersebut memberikan beberapa keuntungan.Pola tersebut diidentifikasi, divalidasi, dan digunakan untuk membuat sebuah prediksi,Witten, Frank, dan Hall[8].

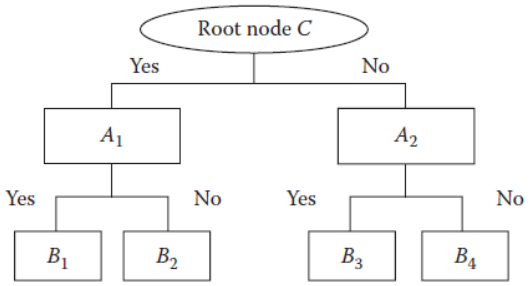
Klasifikasi merupakan salah satu teknik data mining.Klasifikasi (taksonomi) merupakan proses penempatan objek atau konsep tertentu ke dalam satu set kategori berdasarkan objek yang digunakan. Salah satu teknik klasifikasi yang paling popular digunakan adalah *decision tree,* Han dan Kamber[6]. *Decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) di mana setiap *node* merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas.*Node* yang paling atas dari *decision tree* disebut sebagai *root*.

Pada *decision tree* terdapat 3 jenis *node*, yaitu:

a. *Root Node*, merupakan *node* paling atas, pada *node* ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu.
b. *Internal Node* , merupakan *node* percabangan, pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua.
c. *Leaf node atau terminal node* , merupakan *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5, Larose[9]. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record.* Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin, dan temperatur.

Seperti ditunjukkan dalam Gambar 1, *decision tree* tergantung pada aturan *if-then*, tetapi tidak membutuhkan parameter dan metrik. Strukturnya yang sederhana dan dapat ditafsirkan memungkinkan *decision tree* untuk memecahkan masalah atribut *multi-type. Decision tree* juga dapat mengelola nilai-nilai yang hilang atau data *noise,* Dua dan Xian [10].




Gambar 1 Contoh Struktur *Decision Tree*
Sumber: Dua & Xian [10]

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhan algoritma C4.5.Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (*Iterative Dichotomiser*), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning* Han dan Kamber[6].

Serangkaian perbaikan yang dilakukan pada ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan berpengaruh untuk *decision tree* yaitu C4.5. Perbaikan ini meliputi metode untuk menangani *numeric attributes, missing values, noisy* data, dan aturan yang menghasilkan *rules* dari *trees*, Witten, Frank, dan Hall [8].

Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5, Larose [9] yaitu :

1. Mempersiapkan data *training*. Data *training* biasanya diambil dari data *histori* yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy*. Untuk menghitung nilai *entropy* digunakan rumus :

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

Keterangan :
S= Himpunan kasus
n = jumlah partisi S
$P_i$ = proporsi $S_i$ terhadap S
Kemudian hitung nilai *gain* menggunakan rumus :

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

Keterangan :
S = Himpunan Kasus
A = Fitur
n = jumlah partisi atribut A
|Si| = Proporsi Si terhadap S
|S| = jumlah kasus dalam S

3. Ulangi langkah ke 2 dan langkah ke 3 hingga semua *record* terpatisi
4. Proses partisi pohon keputusan akan berhenti saat :
   a. semua*record* dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam*record* yang dipartisi lagi
   c. Tidak ada *record* di dalam cabang yang kosong

## 3. HASIL DAN PEMBAHASAN

**3.1. Analisis Data.** Proses klasifikasi mahasiswa *dropout* menggunakan sembilan langkah dalam KDD (*Knowledge Discovery in Databases*).KDD adalah analisis eksplorasi secara otomatis dan pemodelan *repository* data yang besar. KDD adalah proses terorganisir untuk mengidentifikasi pola dalam data yang besar dan kompleks dimana pola data tersebut ditemukan yang bersifat sah, baru, dan dapat bermanfaat serta dapat dimengerti, Maimon dan Rokach[12]. Adapun Sembilan langkah dalam KDD yaitu:

1. *Developing an understanding of the application domain,* merupakan tahap pemahaman apa yang harus dilakukan dalam penelitian.
2. *Selecting and creating a data set on which discovery will be performed,* memilih dan menciptakan satu set data yang akan digunakan untuk penelitian.
3. *Preprocessing and cleansing,* pada tahap ini kehandalan data ditingkatkan dengan membersihkan data yang tidak lengkap (*missing value*) dan data tidak benar (*noise*). Data hasil *preprocessing* dan *cleansing* diperoleh sejumlah 400 data.
4. *Data Transformation,* Pada tahap ini disusun dan dikembangkan generasi data yang lebih baik untuk data mining. Tahap ini juga merupakan proses transformasi pada data yang telah dipilih sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data. Data akan dipakai untuk penelitian ditransformasikan kedalam kategori seperti pada Appendix 1. Pada tahap ini juga dilakukan pembagian data untuk data training (80%) dan data testing (20%) dengan menggunakan teknik *Systematic Random Sampling*, Sugiana[13] hasilnya 320 data training dan 80 data testing.
5. *Choosing the appropriate Data Mining task*, pada tahap ini memilih teknik data mining yang digunakan yaitu klasifikasi.
6. *Choosing the Data Mining Algorithm,* tahap ini memilih jenis algoritma yang akan digunakan dalam klasifikasi yaitu Algoritma C4.5.
7. *Employing the Data Mining Algorithm,* mengolah data dengan algoritma yang telah dipilih untuk mendapatkan *rule* dari hasil klasifikasi mahasiswa*dropout*.
8. *Evaluation*, evaluasi dilakukan dengan menerapkan pola yang didapat dari proses sebelumnya terhadap data *testing* yang disediakan. Evaluasi dilakukan dengan *confusion matrix* dan kurva ROC.
9. *Using the discovered knowledge, p*ada tahap ini menggunakan pengetahuan yang diperoleh dari proses data mining untuk penerapan pada aplikasi atau lainnya. Pengetahuan klasifikasi mahasiswa potensi *dropout* diterapkan pada data baru untuk membuat klasifikasi mahasiswa yang berpotensi *dropout*.

Hasil klasifikasi mahasiswa dengan Algoritma C4.5 diuji dengan menggunakan *confussion matrix* dan kurva ROC/AUC (*Area Under Cover*), hasilnya:

**3.1.1 *Confusion Matrix.*** Metode ini hanya menggunakan tabel matriks seperti pada Tabel 1, jika dataset hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negative, Bramer[14]. Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy*, *precison*, dan *recall*.*Accuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi, Han dan Kamber [6].Sedangkan *precision* atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya.*Recall* atau *sensitivity* adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar, Powers[15].

Tabel 1. Model *Confusion Matrix*

| *Correct Classification* | *Classified as* | |
|---|---|---|
| | + | - |
| + | *True positives* | *False negatives* |
| - | *False positives* | *True negatives* |

Sumber: Han & Kamber [6]

*True Positive* adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positive* adalah jumlah *record negative* yang diklasifikasikan sebagai positif, *false negative* adalah jumlah *record* positif yang diklasifikasikan sebagai negative, *true negative* adalah jumlah *record negative* yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity (recall), Specifity, precision,* dan*accuracy. Sensitivity* digunakan untuk membandingkan jumlah *t_pos* terhadap jumlah *record* yang positif sedangkan *Specifity, precision* adalah perbandingan jumlah *t_neg* terhadap jumlah *record* yang negative. Untuk menghitung digunakan persamaan dibawah ini, Han dan Kamber[6].

$$\text{Sensitifity} = \frac{t\_pos}{pos}$$

$$\text{Specifity} = \frac{t\_neg}{neg}$$

$$\text{Precision} = \frac{t\_pos}{t\_pos + f\_pos}$$

$$\text{accuracy} = \text{Sensitivity}\frac{pos}{(pos+neg)} + \text{Specifity}\frac{neg}{(pos+neg)}$$

Keterangan :

t_pos : Jumlah*true positives*
t_neg : Jumlah*true negative*
p : Jumlah*record positives*
n : Jumlah*tupel negatives*
f_pos : Jumlah*false positives*

Tabel 2 memperlihatkan tingkat *accuracy*, *precision*, dan *recall* hasil klasifikasi pada data *training* dan *testing* hasil pengujian dengan *confusion matrix.*

Tabel 2 Hasil pengujian dengan *confusion matrix*

| | Data *training* | Data *testing* | Hasil Komparasi |
|---|---|---|---|
| *Accuracy* | 98,75% | 93,75% | 97,75% |
| *Precision* | 92,31% | 62,50% | 86,35% |
| *Recall* | 57,14% | 71,43% | 60% |

**3.1.2 Kurva ROC.** Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual.ROC mengekspresikan *confusion matrix*.ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positive* sebagai garis vertical, Vercellis[16].

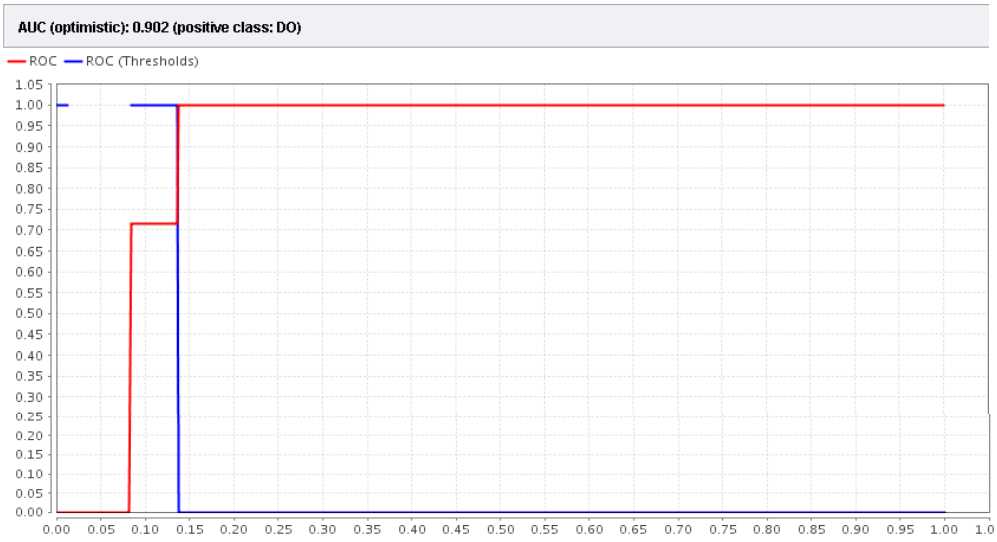
Hasil perhitungan divisualisasikan dengan kurva ROC (*Receiver Operating Characteristic*) atau AUC (*Area Under Curve*). ROC memiliki tingkat nilai diagnosa yaitu, Gorunescu[5]:

a. Akurasi bernilai 0.90 – 1.00 = *excellent classification*
b. Akurasi bernilai 0.80 – 0.90 = *good classification*
c. Akurasi bernilai 0.70 – 0.80 = *fair classification*
d. Akurasi bernilai 0.60 – 0.70 = *poor classification*
e. Akurasi bernilai 0.50 – 0.60 = *failure*

Hasil yang didapat dari pengolahan ROC untuk algoritma C4.5 dengan menggunakan data *training* sebesar 0.999 dapat dilihat pada gambar 2 dengan tingkat diagnosa *excellent classification.*


Gambar 2 Kurva ROC data *training* untuk metode C4.5

Hasil yang didapat dari pengolahan ROC untuk algoritma C4.5 dengan menggunakan data testing sebesar 0.902 dapat dilihat pada gambar 3 dengan tingkat diagnosa *excellent classification.*




Gambar 3 Kurva ROC data *testing* untuk metode C4.5

Klasifikasi mahasiswa yang sudah diuji dengan *confusion matrix* dan kurva ROC diaplikasikan ke dalam data baru untuk pengujian selanjutnya.Hasil pengujian pada data baru menunjukkan tingkat akurasi hasil klasifikasi mahasiswa sebesar 90%. Sehingga *rule* yang diperoleh dari hasil klasifikasi mahasiswa dalam penelitian ini dapat diterapkan ke dalam aplikasi sistem klasifikasi mahasiswa sebagai berikut:


Gambar 4. Program klasifikasi mahasiswa *dropout*

## 3. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian pada klasifikasi mahasiswa potensi *dropout* dapat diambil beberapa kesimpulan sebagai berikut:

1. Klasifikasi mahasiswa dengan algoritma C4.5 dapat mengklasifikasikan mahasiswa aktif dan *dropout*.

2. Hasil evaluasi dan validasi dengan *confussion matrix* menunjukkan tingkat akurasi pada algoritma C4.5 sebesar 97,75%.
3. Hasil evaluasi dan validasi dengan ROC/AUC menunjukkan nilai lebih dari 0,9 sehingga dapat dimasukkan kedalam *excellent classification*.
4. Penerapan *rule* dari algoritma C4.5 yang digunakan dalam klasifikasi mahasiswa potensi *dropout* terhadap data baru diperoleh hasil evaluasi dan validasi dengan *confussion matrix* menghasilkan tingkat akurasi sebesar 90,00%.

Saran yang diajukan dalam penelitian ini yaitu untuk penelitian selanjutnya dengan permasalahan yang sama dan dengan metode yang sama dapat ditingkatkan salah satunya dengan melakukan *pruning* terhadap algoritma C4.5 sehingga pohon yang terbentuk tidak terlalu besar bahkan mungkin untuk jumlah data yang besar sekalipun. Ini dilakukan untuk mengefisienkan kinerja dari algoritma C4.5 tanpa mengurangi keakuratannnya.

## 4. DAFTAR REFERENSI


[1] Harahap, S. (2006). *Penegakan Moral Akademik Didalam dan Luar Kampus.* Jakarta: Raja Grafindo.

[2] Baharudin & Makin, M. (2004). *Pendidikan: Suatu Pendekatan Praktek*. Jakarta: AR-RUZZ Media.

[3] Sobur, A. (2006). *Psikologi Umum*. Bandung: Pustaka Setia.

[4] Deker, G. W., Pechenizkiy, M., Vleeshouwers, J. M,. (2009). Predicting Students Drop Out: A Case Study.

[5] Gorunescu, F. (2011). *Data Mining Concept Model and Techniques*. Berlin: Springer. ISBN 978-3-642-19720-8

[6] Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques.* San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3

[7] Guidici, P. & Figini, S. (2009*). Applied Data Mining for Business and Industry* (2nd ed). Italy. John Wiley & Sons, Ltd. ISBN: 978-0-470-05886-2

[8] Witten, I. H., Frank, E., Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (3rd ed). USA: Elsevier

[9] Larose, D. T. (2005). *Discovering Knowledge in Data.* New Jersey: John Willey & Sons, Inc. ISBN 0-471-66657-2.

[10] Dua, S. & Xian Du. (2011). *Data Mining and Machine Learning in Cybersecurity*. USA: Taylor & Francis Group. ISBN-13: 978-1-4398-3943-0

[11] C.R.Kothari. (2004). *Research Methology Methods and Techniques.* India: New Age International Limited. ISBN (13) : 978-81-224-2488-1

[12] Maimon, Oded.,& Rokach, Lior. (2010). *Data Mining and Knowledge Discovery Handbook, 2nd Edition*. New York: Springer. ISBN 978-0-387-09822-7

[13] Sugiana, Dadang. (2008). Secuil Tentang Sampling dalam Penelitian Kuantitatif**.** Juli 18, 2012. http ://dankfsugiana.wordpress.com/2008/07/08/ populasi-dan-teknik-sampling /

[14] Bramer, Max. (2007). *Principles of Data Mining*. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

[15] Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation. *Journal of Machine Learning Technologies*,ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63

[16] Vercellis, Carlo. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. United Kingdom: John Willey & Son.

Appendix 1

| Atribut | Nilai |
|---|---|
| waktu kuliah | pagi |
| | malam |
| ipk smt 1 | >=3,00 |
| | 2,00-2,99 |
| | 1,00-1,99 |
| | <1,00 |
| kehadiran smt 1 | 75%-100% |
| | 50%-74% |
| | 25%-49% |
| | 0%-24% |
| jenis kelamin | L |
| | P |
| usia masuk | <=25 |
| | >25 |
| asal daerah | dalam propinsi |
| | luar propinsi |
| | luar jawa |
| jurusan SLTA | MIPA |
| | ilmu sosial |
| | ilmu komputer |
| | bahasa |
| | lain-lain |
| orangtua | Ada |
| | yatim |
| penghasilan ortu/wali | sangat rendah |
| | rendah |
| | sedang |
| | tinggi |
| biaya studi | orangtua/wali |
| | sendiri |
| | beasiswa |
| bekerja | Ya |
| | tidak |
| beasiswa | beasiswa |
| | tidak |