Analisis Algoritma Decision Tree untuk Prediksi Mahasiswa Non Aktif

Khafiizh Hastuti[1], Erwin Yudi Hidayat[2]
[1, 2] *Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang 50131*
E-mail : [1] afis@dsn.dinus.ac.id, [2] erwin@dsn.dinus.ac.id

ABSTRAK

Presentasi mahasiswa yang lulus tepat waktu pada perguruan tinggi merupakan salah satu faktor penentu kualitas perguruan tinggi. Data dari Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia menunjukkan bahwa presentasi kelulusan mahasiswa tepat waktu hanya mencapai 51,97% saja dan 48,03% tidak diketahui statusnya apakah menempuh studi tidak tepat waktu, memiliki status non aktif, atau drop out. Perguruan tinggi memiliki data akademik dan biodata mahasiswa sejak mereka mendaftar hingga lulus kuliah. Algoritma klasifikasi data mining Decision Tree dapat digunakan untuk melalukan prediksi mahasiswa non aktif, sehingga dapat memberikan warning bagi mahasiswa untuk lebih meningkatkan prestasi belajar dan mencegah kasus mahasiswa drop out.

Kata kunci : *decision tree,* mahasiswa non aktif

1. PENDAHULUAN

Presentasi mahasiswa yang lulus tepat waktu pada perguruan tinggi merupakan salah satu faktor penentu kualitas perguruan tinggi. Berdasarkan matriks penilaian instrument akreditasi program studi Badan Akreditasi Nasional Perguruan Tinggi [1] bahwa persentase mahasiswa yang lulus tepat waktu merupakan salah satu elemen penilaian akreditasi universitas. Data yang didapatkan dari Data dari Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia [2] pada tahun akademik 2001/2002 sampai dengan 2009/2010 menunjukkan bahwa perguruan tinggi menerima rata-rata sebanyak 868.050 mahasiswa baru dan meluluskan rata-rata 451.168 mahasiswa setiap tahunnya atau hanya mencapai 51,97% saja. Dari data tersebut diketahui bahwa 48,03% mahasiswa tidak diketahui statusnya, apakah mahasiswa menempuh studi tidak tepat waktu, memiliki status non aktif, atau drop out.

Perguruan tinggi perlu mengetahui faktor-faktor penyebab kegagalan studi. Database perguruan tinggi memiliki data akademik dan biodata mahasiswa sejak mereka mendaftar kuliah sampai dengan lulus. Data tersebut apabila digali dengan tepat dapat digunakan untuk mendapatkan pengetahuan atau pola untuk pengambilan keputusan [3]. Serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan *data mining* [4]. Kotsiantis, Pierrakeas dan Pintelas [5] menyebutkan bahwa sangat penting bagi dosen untuk mendeteksi mahasiswa yang cenderung *drop out* sebelum mereka memasuki pertengahan masa studi.
Data mining dalam dunia pendidikan dikenal dengan istilah *Educational Data Mining* (EDM) [6]. EDM dapat membantu pendidik untuk menganalisis cara belajar, mendeteksi mahasiswa yang memerlukan dukungan, dan memprediksi kinerja siswa. Beberapa algoritma klasifikasi data mining dapat digunakan untuk mencegah secara dini kegagalan mahasiswa. Penelitian yang dilakukan oleh Gerben W. Dekker [7] menyebutkan bahwa monitoring dan dukungan terhadap mahasiswa di tahun pertama sangat penting dilakukan. Mahasiswa jurusan teknik elektro Universitas Eindhoven yang berhenti studi pada tahun pertama mencapai hingga 40%. Kurikulum yang sulit dianggap sebagai salah satu penyebab tingginya jumlah mahasiswa *drop out.* Selain itu, nilai, prestasi, kepribadian,latar belakang sosial mempunyai peran dalam kesuksesan akademik mahasiswa. Dekker menggunakan algoritma *Decision tree, Bayesian classifiers, logistic models, rule-based learner* dan *random forest.*

Penelitian ini menggunakan algoritma *decision tree* untuk melakukan prediksi mahasiswa non aktif dengan menggunakan data yang ada pada Universitas Dian Nuswantoro dengan menggunakan 3681 data set mahasiswa yang terdiri atas data demografi dan akademik mahasiswa.

2. TINJAUAN PUSTAKA

Menurut Witten [4], proses untuk mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining. Sotiris Kotsiantis [8] mengelompokkan 354 mahasiswa Hellenic Open University, menjadi 2 kelompok atribut yaitu berbasis kurikulum dan kinerja mahasiswa. Atribut kelompok berbasis kurikulum terdiri atas jenis kelamin, usia, status marital, jumlah anak,

pekerjaan, kemampuan komputer, hubungan pekerjaan dengan komputer. Atribut kelompok kinerja mahasiswa terdiri atas tatap muka ke-1, tugas ke-1, tatap muka ke-2, tugas ke-2.

2.1 Decision Tree

Decision tree digunakan untuk kasus yang memiliki ciri-ciri sebagai berikut [9]:

1. Data atau contoh dinyatakan dengan pasangan atribut dan nilainya.
2. Label atau output data biasanya bernilai diskrit.
3. Data mempunyai *missing value*

Teori entropi diadopsi untuk memilih pemecahan atribut yang tepat untuk algoritma C4.5, dengan menyatakan jumlah rata-rata informasi yang dibutuhkan untuk mengklasifikasikan sampel.
Untuk menghitung nilai entropy digunakan rumus:

$$entropy(S) = \sum_{i=1}^{n} -p_i . log_2 p_i \quad (1)$$

Dimana *S* merupakan himpunan kasus, *n* adalah jumlah partisi *S*, dan p_i adalah proporsi S_i terhadap S.Ketika output data atau variabel dependent *S* dikelompokkan berdasarkan atribut *A*, dinotasikan dengan $gain(S, A)$. hasil dari atribut mendapatkan *information gain* yang didefinisikan sebagai:

$$gain(S, A) = entropy(S) - \sum_{i=}^{n} \frac{S_i}{S} . entropy(S_i) \quad (2)$$

Dimana *S* merupakan himpunan kasus, *A* adalah atribut, *n* adalah jumlah partisi atribut A, S_i adalah proporsi S_i terhadap *S* dan S adalah jumlah kasus dalam himpunan. Sebuah prosedur tambahan dilakukan untuk menghindari pohon yang menghasilkan *overfits* data yang kompleks.
Berikut ini adalah langkah-langkah algoritma *decision tree*:

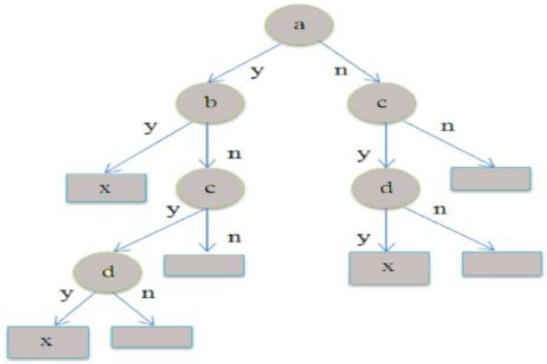
1. Input data.
2. Hitung konsep entropi:

$$entropy(S) = -p_1 . log_2 p_1 - p_2 . log_2 p_2 \dots p_n . log_n p_n$$

3. Hitung information gain.

$$gain(S, A) = entropy(S) - \sum_{i=}^{n} \frac{S_i}{S} . entropy(S_i)$$

Untuk memudahkan pembacaan pohon keputusan dibuatlah set aturan if-then, dimana satu aturan digeneralisasikan ke seluruh simpul daun. Sebuah ilustrasi dijelaskan pada Gambar 1, dimana sebuah aturan (rule) dengan struktur yang sama tetapi berbeda atribut, seperti:



Gambar 1. Decision Tree untuk Disjungsi Sederhana

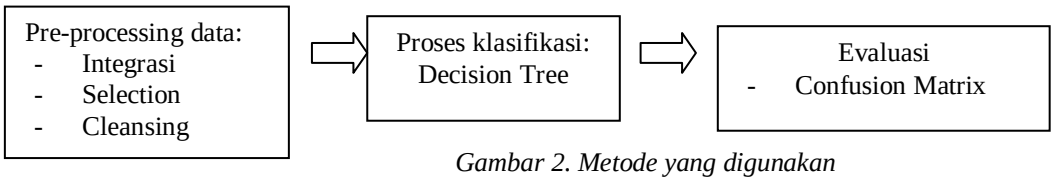
2.2 Mahasiswa Non Aktif

Mahasiswa non aktif adalah mahasiswa yang tidak melakukan registrasi administratif setiap awal semester gasal [10]. Mahasiswa yang memiliki status non aktif selama empat semester berturut-turut dikategorikan sebagai mahasiswa drop out.

3. METODE PENELITIAN

Tahapan penelitian yang dilakukan adalah dengan melakukan pengumpulan data, pengolahan awal data, metode yang digunakan, eksperimen dan pengujian model, evaluasi dan validasi hasil klasifikasi.

1. Pengumpulan data
   Jumlah mahasiswa keseluruhan adalah 13.416 mahasiswa terdiri atas lima fakultas baik jenjang strata satu maupun diploma tiga dari angkatan 1993 sampai dengan angkatan 2011. Dari jumlah tersebut, tercatat terdapat 4.138 diantaranya memiliki status non aktif dan 9.278 mahasiswa memiliki status aktif. Sampel data demografi mahasiswa yang terdiri atas NIM, nama, tempat lahir, tanggal lahir, alamat di kota Semarang, alamat asal, kota asal dan seterusnya. sampel data akademik mahasiswa yang terdiri atas Nim, nama, Indek Prestasi Semester 1, Indek Prestasi Semester 2, Indek Prestasi Semester 3, Indek Prestasi Semester 4, jumlah SKS (Satuan Kredit Semester) yang diambil mahasiswa tiap semester ganjil dan genap.
2. Pengolahan awal data
   - Data integrasi
     Data mahasiswa yang terdiri atas data akademik dan data demografi yang diperoleh dintegrasikan sebagai satu kesatuan data.
   - Seleksi fitur (atribut)
     Seleksi fitur digunakan sebagai input untuk proses klasifikasi. Seleksi fitur dilakukan dengan mengambil sebagian variabel pada seluruh atribut yang ada pada data untuk dijadikan atribut penentu dalam melakukan pemberian keputusan. Fitur yang digunakan adalah sebagai berikut: program studi, jenis kelamin, usia saat mendaftar, kota asal, status domisili, agama, marital, asal sekolah, status kerja, asal biaya, pekerjaan orang tua, penghasilan orang tua, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, SKS 1, SKS 2, SKS 3, SKS 4, status skripsi, status akademik.
   - Data cleansing
     Pada tahap ini, dilakukan penghapusan data yang tidak lengkap. Penulis menggunakan tiga program studi jenjang strata satu pada Fakultas Ilmu Komputer sebagai data set. Dari jumlah 13.416 mahasiswa, data yang layak digunakan sebanyak 3.861 mahasiswa dari program studi Teknik Informatika, Sistem Informasi dan Desain Komunikasi Visual jenjang strata satu angkatan 2005 sampai dengan 2009. Tercatat 1.018 mahasiswa memiliki status non aktif dan 2.843 mahasiswa dengan status aktif
3. Metode yang digunakan
   Proses secara bertahap dimulai dari pengolahan data *pre-processing data* yaitu *integrasi, selection* dan *cleansing.* Algoritma yang digunakan adalah *decision tree*.



Gambar 2. Metode yang digunakan

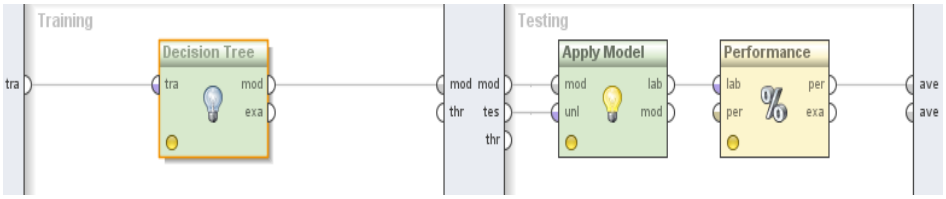
Setelah hasil proses klasifikasi diketahui, selanjutnya evaluasi dengan menggunakan *confusion matrix,* untuk mengetahui akurasi algoritma *decision tree* untuk prediksi mahasiswa non aktif.

4. Eksperimen dan Pengujian Model

Pada tahap ini dilakukan eksperimen dan teknik pengujian yang digunakan untuk mengukur tingkat akurasi algoritma berdasarkan data set mahasiswa yang digunakan yaitu sebanyak 3.861 mahasiswa dari program studi Teknik Informatika, Sistem Informasi dan Desain Komunikasi Visual jenjang strata satu tahun akademik 2005 sampai dengan 2009. Tercatat 1.018 mahasiswa memiliki status non aktif dan 2.843 mahasiswa dengan status aktif.

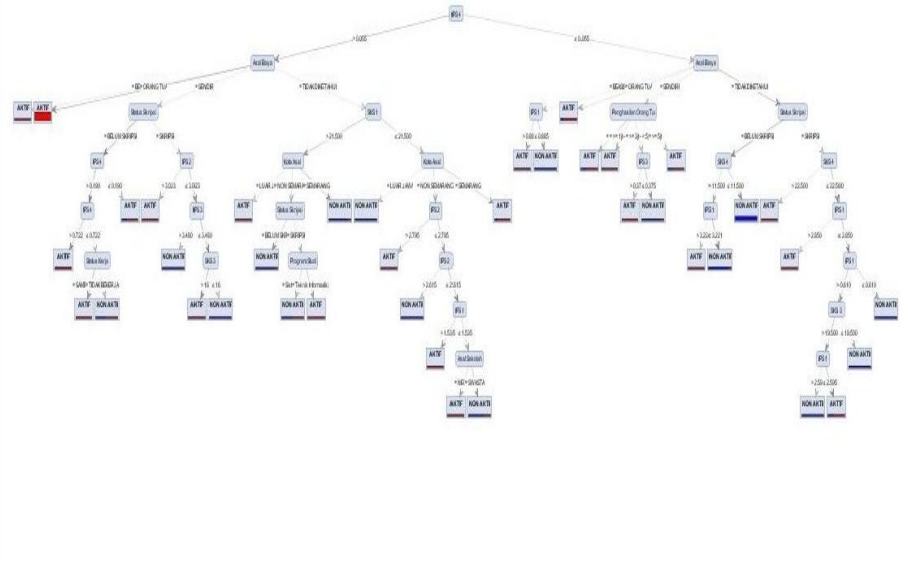
Analisis Pengujian Menggunakan Decision Tree

Proses evaluasi algoritma *decision tree* dengan menggunakan Rapid Miner ditunjukkan dalam desain model seperti yang terlihat dalam Gambar 3 di bawah ini:



Gambar 3. Desain model *decision tree*

Algoritma yang digunakan adalah algoritma C4.5, dengan membentuk pohon keputusan. Proses ini menghitung jumlah mahasiswa aktif dan mahasiswa non aktif beserta entropy dari semua kasus berdasarkan atribut program studi, jenis kelamin, usia saat mendaftar, kota asal, status domisili, agama, marital, asal sekolah, status kerja, asal biaya, pekerjaan orang tua, penghasilan orang tua, IPS1, IPS2, IPS3, IPS4, SKS1, SKS2, SKS3, SKS4 dan status skripsi.
Hasil dari algoritma C4.5 menghasilkan pohon keputusan seperti Gambar 4 di bawah ini:



Gambar 4. Pohon Keputusan C4.5

Tujuan dari menganalisis data dengan menggunakan algoritma *decision tree* adalah ingin mendapatkan *rule* [4] yang akan dimanfaatkan untuk pengambilan keputusan pada data baru. Aturan-aturan yang muncul adalah sebagai berikut:

1. *R1: if ips4>0.055* AND *asal biaya=sendiri* AND *status skripsi=belum skripsi* AND *ips4<=0.722* AND*>0.190* AND *status kerja=tidak bekerja* THEN *status akademik=non aktif*
2. R2: *if ips4>0.055* AND *asal biaya=sendiri* AND *status skripsi=skripsi* AND *ips2<=3.023* AND *ips3 >3.480* THEN *status akademik=non aktif*
3. R3: *if ips4>0.055* AND *asal biaya=sendiri* AND *status skripsi=skripsi* AND *ips2<=3.023* AND *ips3<=3.480* AND *sks3<=16* THEN *status akademik=non aktif*
4. R4: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1>21.5* AND *kota asal=non semarang* AND *status skripsi=belum skripsi* THEN *status akademik=non aktif*
5. R5: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1>21.5* AND *kota asal=non semarang* AND *status skripsi=skripsi* AND *program studi=sistem informasi* THEN *status akademik=non aktif*
6. R6: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1>21.5* AND *kota asal= semarang* THEN *status akademik=non aktif*
7. R7: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1<=21.5* AND *kota asal= luar jawa* THEN *status akademik=non aktif*

8. R8: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1<=21.5* AND *kota asal=non semarang* AND *ips2>2.615* AND*<=2.795* AND *ips1<=1.535 and asal sekolah=swasta* THEN *status akademik=non aktif*
9. R9: *if ips4>0.055* AND *asal biaya=tidak diketahui* AND *sks1<=21.5* AND *kota asal=non semarang* AND *ips2>2.615* AND *<=2.795* THEN *status akademik=non aktif*
10. R10: *if ips4<=0.055* AND *asal biaya=beasiswa* AND *ips1<=0.885* THEN *status akademik=non aktif*
11. R11: *if ips4<=0.055* AND *asal biaya=sendiri* AND *penghasilan orang tua>=3jt-<5jt* AND *ips3<=0.375* THEN *status akademik=non aktif*
12. R12: *if ips4<=0.055* AND *asal biaya=tidak diketahui* AND *status skripsi=belum skripsi* AND *sks4>11.5* AND *ips1<=3.221* THEN *status akademik=non aktif*
13. R13: *if ips4<=0.055* AND *asal biaya=tidak diketahui* AND *status skripsi=belum skripsi* AND *sks4<=11.5* THEN *status akademik=non aktif*
14. *R14: if ips4<=0.055* AND *asal biaya=tidak diketahui* AND *status skripsi= skripsi* AND *sks4<=22.5* AND *ips1>0.61*AND*<=2.850* AND *sks3>19.5* AND *ips1>2.595* THEN *status akademik=non aktif*
15. R15: *if ips4<=0.055* AND *asal biaya=tidak diketahui* AND *status skripsi= skripsi* AND *sks4<=22.5* AND *ips1>0.61*AND*<=2.850* AND *sks3>19.5* THEN *status akademik=non aktif*
16. R16: *if ips4<=0.055* AND *asal biaya=tidak diketahui* AND *status skripsi= skripsi* AND *sks4<=22.5* AND *ips1<=2.85* AND *ips1<=0.61* THEN *status akademik=non aktif*

5. Evaluasi dan Validasi Hasil

Data yang telah diuji kemudian dievaluasi dalam *confusion matrix* untuk diketahui akurasi, presisi dan *type error*. Pengukuran akurasi dengan menggunakan decision tree dapat dilihat pada Gambar 5 di bawah ini:

accuracy: 95.29% +/- 1.29% (mikro: 95.29%)

	true NON AKTIF	true AKTIF	class precision
pred. NON AKTIF	921	85	91.55%
pred. AKTIF	97	2758	96.60%
class recall	90.47%	97.01%	

Gambar 5. Akurasi prediksi mahasiswa non aktif menggunakan *decision tree*

Diketahui:

True Positive = 921
False Positif = 85
False Negative = 97
True Negative = 2758

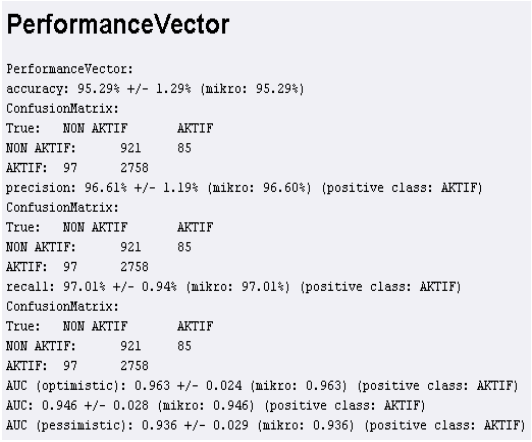
Dihitung:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (921 + 2758) / (921 + 2758 + 85 + 97)
= 3679 / 3861
= 0,9529

Precision positive = TP / (TP + FP)
= 921 / (921 + 85)
= 921 / 1006
= 0,9155

Precision negative = TN / (TN + FN)
= 2758 / (2758 + 97)
= 2758 / 2855
= 0,9660

Type error = (FN) / (TP + FP + TN + FN)
= (97) / (921 + 85 + 2758 + 97)
= 971/ 3861
= 0,2515



Gambar 6. *Performance vector decision tree*

Pada Gambar 6 menunjukkan hasil pengujian dengan menggunakan model *decision tree*, tingkat akurasi mencapai 95,29%.

4. KESIMPULAN

Decision tree menghasilkan akurasi yang sangat baik yaitu mencapai 95,29% untuk mempredikasi mahasiswa non aktif.
Saran penelitian selanjutnya dengan menambahkan atribut seperti nilai ujian nasional calon mahasiswa, prestasi siswa di sekolah asal, jalur masuk pendaftaran siswa, dan lain-lain.

DAFTAR PUSTAKA

[1] Buku VI Matriks Penilaian Instrumen Akreditasi Program Studi Badan Akreditasi Nasional Perguruan Tinggi, 2008.

[2] Pusat Statistik Pendidikan Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional Republik Indonesia.

[3] Alaa El-Halees, "Department of Computer Science," Mining Students Data to Analyze Learning Behaviour: A Case Study, Journal of Educational Data Mining,2009.

[4] Ian H. Witten, frank Eibe, and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Asma Stephan and Burlington, Eds. United States of America: Morgan Kaufmann, 2011.

[5] S.B. Kotsiantis, C.J. Pierrakeas, and P.E. Pintelas, "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques," *In International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Oxford*, 3-5, 2003.

[6] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting School Failure Using Data Mining," Journal of Educational Data Mining, 2011.

[7] Gerben W. Dekker, "Predicting Students Drop Out: A Case Study," In International Conference on Educational Data Mining, Cordoba, Spain, 41-50, 2009.

[8] Sotiris Kotsiantis, "Educational Data Mining: A Case Study for Predicting Dropout-Prone Students," Int. J. of Knowledge Engineering and Soft Data Paradigms, vol. X, 2010.

[9] Budi Santoso, Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis, 1st ed. Yogyakarta, Indonesia: Graha Ilmu, 2007.

[10] Keputusan Rektor Universitas Dian Nuswantoro nomor: 075/KEP/UDN-01/IV/2009 tentang Peraturan Akademik Universitas Dian Nuswantoro tahun akademik 2009/2010.