IMPLEMENTASI ALGORITMA ID3 UNTUK KLASIFIKASI PERFORMANSI MAHASISWA (STUDI KASUS ST3 TELKOM PURWOKERTO)

Andika Elok Amalia[1)], Muhammad Zidny Naf'an[2)]

[1), 2)] Program Studi Informatika ST3 Telkom
Jl D.I. Panjaitan no. 128, Purwokerto53147
Email : andika.amalia@st3telkom.ac.id[1)], zidny@st3telkom.ac.id[2)]

Abstrak

Prediksi performansi mahasiswa sangat penting untuk diketahui agar dapat dilakukan langkah preventif terkait hasil akademik yang buruk. Dalam penelitian ini dimanfaatkan data mahasiswa yang telah melewati perkuliahan selama satu tahun atau dua semester awal pada Sekolah Tinggi Teknologi Telematika Telkom. Data-data tersebut kemudian digali menggunakan metode klasifikasi data mining dengan algoritma ID3. Sebelum dilakukan penggalian, data sudah terlebih dahulu dipilih beberapa variabel yang mempengaruhi performansi dan dibersihkan. Hasil dari penelitian ini adalah sebuah pohon keputusan (decision tree) untuk mengklasifikasi mahasiswa baru tersebut dan dimanfaatkan untuk menentukan, mahasiswa yang perlu mendapatkan matrikulasi.
Kata kunci: *data mining, ID3, decision tree, prediksi, klasifikasi.*

1. Pendahuluan

Pendidikan yang berkualitas pada perguruan tinggi tidak hanya ditentukan oleh cara pengajaran namun juga penanganan terhadap latar belakang dari mahasiswa yang ada pada perguruan tinggi tersebut. Performansi akademik mahasiswa yang dapat diprediksi sebelumnya akan menjadi alat pendukung untuk memetakan mahasiswa. Pemetaan tersebut dapat memunculkan strategi untuk menindaklanjuti mahasiswa yang diprediksi memiliki performansi kurang baik. Langkah preventif dapat dilakukan sehingga kompetensi pengajaran dapat dicapai oleh seluruh mahasiswa tanpa kecuali. Seluruh perguruan tinggi pasti memiliki data awal calon mahasiswa yang dapat dimanfaatkan untuk mencari pola klasifikasi mahasiswa.

Beberapa permasalahan yang terjadi pada ST3 Telkom adalah kesulitan dosen dalam mengajar di kelas karena kemampuan mahasiswa dalam satu kelas beraneka ragam dan kondisi hasil belajar mahasiswa dilihat dari indeks prestasi kumulatif setelah tahun pertama yang menunjukkan *gap* sangat besar. Oleh karena itu, diperlukan langkah preventif dengan mengadakan kelas matrikulasi. Namun untuk menentukan mahasiswa yang harus mengikuti program tersebut diperlukan teknik agar tepat sasaran dan berjalan efisien karena tidak semua mahasiswa baru harus mengikutinya.

Data-data mahasiswa khususnya nilai pada saat sekolah menengah biasanya hanya digunakan untuk seleksi masuk pada perguruan tinggi setelah itu tidak dimanfaatkan lagi. Padahal data-data yang tidak terpakai tersebut dapat digunakan untuk menggali informasi lebih dalam, salah satunya untuk memprediksi performansi mahasiswa menggunakan teknik *data mining*. Menurut Han dan Kamber (Han and Kamber, 20016), *data mining* adalah proses penggalian pengetahuan dari data dalam jumlah besar [1].

Banyak algoritma dalam *data mining* yang dapat dilakukan untuk klasifikasi. Pembahasan mengenai sistem prediksi terhadap level kelulusan mahasiswa pada sebuah universitas sudah pernah dilakukan (Ogunde dan Ajibade , 2014) dengan menggunakan Algoritma Decision Tree ID3 dengan variabel yang diolah nilai-nilai mahasiswa saat masih pada tingkat sekolah menengah. Pada penelitian tersbut mahasiswa diklasifikasi dalam 5 kelas yaitu lulus dengan *grade first class*, *second class upper*, *second class lower*, *third class* dan *pass* (tidak lulus). Hasil dari penelitian tersebut berupa model pohon keputusan dengan *true positive rate* bagus dan dapat dimanfaatkan oleh manajemen universitas serta perencana studi mahasiswa. Saran untuk pekerjaan serupa adalah menggunakan metode lain dari *decision tree* pada *data mining* [2].

Studi yang sama untuk memprediksi performansi akademis mahasiswa dengan variabel tidak hanya nilai saat sekolah menengah, namun juga mempertimbangkan varibel lain di luar nilai seperti jenis kelamin, penghasilan orang tua dan lain sebagainya juga sudah pernah dilakukan (Pal, 2013). Pada penelitian ini, proses klasifikasi menggunakan lebih dari satu metode pada *decision tree* yaitu ID3, C4.5 dan Bagging. Penelitian tersebut menghasilkan kesimpulan bahwa algoritma *decision tree* ID3 yang memiliki tingkat akurasi yang paling baik dan algoritma C4.5 dapat belajar efektif untuk membuat model pada kasus prediksi nilai mahasiswa yang diakumulasi dari tahun-tahun sebelumnya [3].

Metode lain yang telah diterapkan pada kasus sama adalah Naïve Bayes (Bhardwaj , 2011). Pada peneltian tersebut variabel yang digunakan tidak jauh berbeda dari yang studi yang dilakukan pada studi setelahnya [2]. Dari penelitian ini disimpulkan bahwa ada 3 varibel teratas yang memiliki potensi paling berpengaruh

terhadap performansi mahasiswa yaitu nilai ujian pada sekolah menengah, lokasi tempat tinggal dan bahasa yang digunakan sebagai media pembelajaran[4].
Berdasarkan penelitian-penelitian di atas, algoritma ID3 memberikan hasil paling baik. Dalam penelitian ini, digunakan teknik *data mining* dengan menerapkan algoritma ID3 pada data awal calon mahasiswa dan klasifikasi IPK mahasiswa. Hasil dari penelitian ini berupa pohon keputusan atau sekumpulan aturan yang dapat digunakan untuk memprediksi kelas calon mahasiswa setelah setahun perkuliahan.

2. Pembahasan

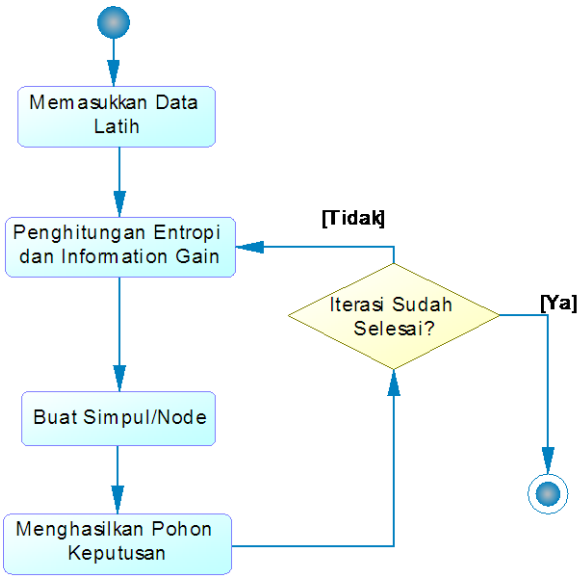
2.1 Algoritma ID3

ID3 merupakan salah satu pendekatan klasifikasi dalam data mining dengan menciptakan pohon berdasarkan atribut yang ada untuk mengatasi suatu permasalahan.. Pohon keputusan adalah sebuah pohon dimana masing-masing cabang dari simpul merepresetasikan alternatif pilihan dan masing-masing ujung simpul/*node* merepresentasikan keputusan. ID3 mengambil konsep dari teori informasi dimana pemilihan atribut untuk membentuk pohon tersebut dilakukan dengan properti statistik yang disebut *information gain*. Nilai *gain* digunakan untuk mengukur kualitas suatu atribut dalam memisahkan training example ke dalam kelas target. [5]
Dalam penentuan akar *(root)* dan cabang, ID3 menggunakan nilai *information gain* terbesar dari atribut-atribut yang ada. *Information gain* diperoleh dari rumus [6]:

$$Gain(S,F) = Entropy(S) - \sum_{f \in values(F)} \frac{|S_f|}{|S|} Entropy(S_f) \qquad (1)$$

Entropi adalah jumlah dari informasi yang terdapat pada atribut yang diperoleh dari rumus [5] :

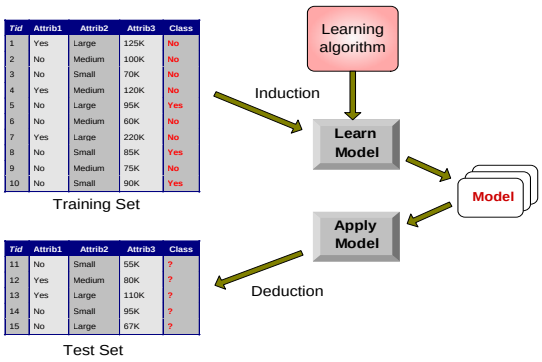
$$Entropy(S) = \sum_{I=1}^{C} - P_i \log_a Pi \qquad (2)$$



Gambar 1. *Diagram Alir Algoritma ID3*

2.2 Proses Data Mining

Metode pada penelitian terdiri dari 3 langkah yaitu pengumpulan data serta penentuan variabel, *pre processing* dan proses penggalian data. Secara umum model klasifikasi data mining dapat dilihat pada Gambar 1.



Gambar 2. *Model Klasifikasi Data Mining*

1) Pengumpulan Data dan Penentuan Variabel
Pada penelitian ini digunakan data mahasiswa yang diterima di ST3 Telkom pada tahun 2015 Program Studi S1 Informatika yang telah setahun melaksanakan perkuliahan dan sudah memiliki variabel kriteria mahasiswa. Total sampel yang didapatkan untuk dijadikan data latih sebanyak 100 data. Terdapat 5 variabel yang diperkirakan mempengaruhi performansi setahun pertama perkuliahan. Variabel nilai-nilai ujian mata pelajaran yaitu matematika, fisika dan Bahasa Inggris dipilih dengan melihat pada jenis mata kuliah yang akan diambil oleh mahasiswa pada tahun pertama. Selain itu jurusan asal mahasiswa serta jeda mahasiswa setelah lulus sekolah menengah atas juga sangat berpengaruh pada proses belajar.

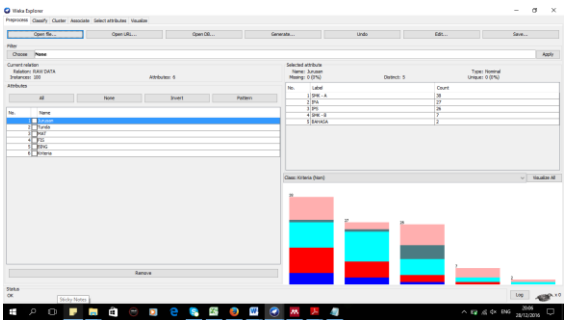
Tabel 1.*Variabel Terkait*

Variabel	Deskripsi	Nilai
Jurusan	Jurusan saat sekolah menengah atas	{ IPA, IPS, SMK A (SMK yang jurusannya terkait dengan program studi), SMK B (SMK yang jurusannya tidak terkait dengan program studi)}
Tunda	Menyatakan adanya waktu penundaan sebelum melanjutkan stud	{YA, TIDAK}
MAT	Nilai rata-	{ A1

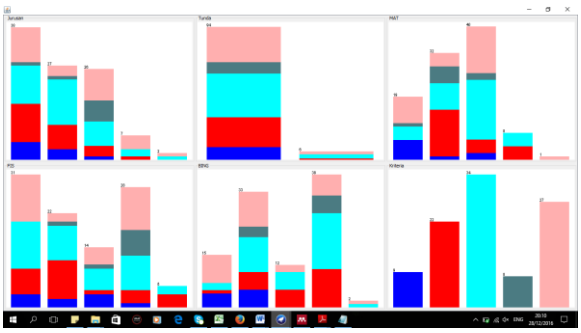
	rata sekolah untuk mata pelajaran matematika	jika >= 85, B2 jika 80 – 84, C3 jika 75 – 79, D4 jika 70 – 74, E6 jika 65 – 69, F7 jika 60 – 64}
BING	Nilai rata-rata sekolah untuk mata pelajaran bahasa inggris	{ A1 jika >= 85, B2 jika 80 – 84, C3 jika 75 – 79, D4 jika 70 – 74, E6 jika 65 – 69, F7 jika 60 – 64}
FIS	Nilai rata-rata sekolah untuk mata pelajaran fisika	{ A1 jika >= 85, B2 jika 80 – 84, C3 jika 75 – 79, D4 jika 70 – 74, E6 jika 65 – 69, F7 jika 60 – 64}
Kriteria	Klasifikasi mahasiswa berdasarkan IPK	{A jika 3,50 – 4,00, B jika 3,00 – 3,49, C jika 2,50 – 2,99, D jika 2.00 - 2.49, E jika <=1.99 }

2) Pre Processing

Pre processing merupakan proses pengolahan data sebelum diterapkan algoritma ID3. Pada tahap ini dilakukan pembersihan data untuk menghilangkan data yang tidak memiliki nilai dan juga *missing value* [7]. Pada tahap ini dan tahap selanjutnya digunakan *tools* yaitu Wakaito Environment for Knowledge Analysis (WEKA).



Gambar 3. *Antarmuka untuk Pre Processing*



Gambar 4. *Visualisasi per Atribut*

Jumlah data per kriteria dapat dijelaskan dengan tabel berikut.

Tabel 2. *Jumlah data per Kriteria*

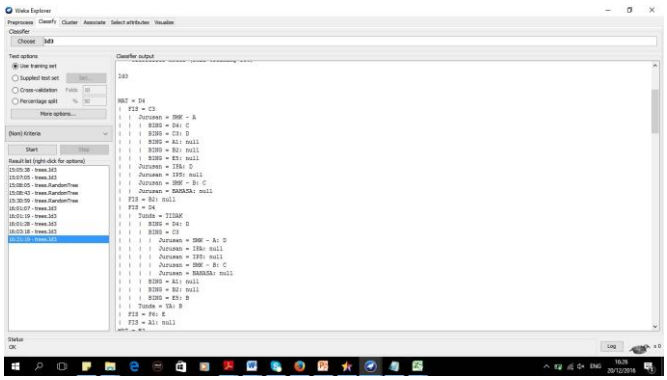
Kriteria	Jumlah
A	22
B	34
C	27
D	9
E	8

3) *Implementasi Data Mining*

Tahap ini merupakan inti dari penelitian yaitu penerapan *data mining* pada data yang sebelumnya sudah diolah. Algoritma ID3 digunakan untuk mengklasifikasi data-data tersebut.

2.3 Hasil dan Pembahasan

Proses *data mining* menghasilkan sekumpulan aturan untuk menentukan kelas mahasiswa setelah melaksanakan setahun perkuliahan.



Gambar 5. *Kumpulan Rule yang Dihasilkan*

Evaluasi terhadap model dilakukan dengan menggunakan keseluruhan data latih yang menghasilkan *confussion matrix* seperti pada Gambar 6, sedangkan performansi yang dihasilkan disajikan pada Tabel 4 dan 5.

```
a  b  c  d  e   <-- classified as
8  0  0  0  1 |  a = D
0 19  3  0  0 |  b = A
1  2 30  0  1 |  c = B
0  4  1  1  2 |  d = E
4  2  7  0 14 |  e = C
```

Gambar 6. *Confussion Matrix*

Tabel 3. *Penjelasan Confussion Matrix*

Baris (Setelah Label)	Penjelasan
1	Nilai {8 0 0 0 1} menunjukkan bahwa ada (8+0+0+0+1) *instances* kelas D yang 8 diklasifikasikan benar dan 1 salah diklasifikasikan sebagai kelas C.
2	Nilai {0 19 3 0 0} menunjukkan ada (0+19+3+0+0) *instances* kelas A, 19 diantaranya diklasifikasikan benar dan 3 diantaranya diklasifikasikan sebagai kelas B.
3	Nilai {1 2 30 0 1} menunjukkan ada (1+2+30+0+1) instances kelas B, 30 diantaranya diklasifikasikan benar, 1sebagai kelas D, 2 diklasifikasikan sebagai kelas A dan 1 diklasifikasikan sebagai kelas C.
4	Nilai {0 4 1 1 2} menunjukkan ada (0+4+1+1+2) instances kelas E, 1 diantaranya diklasifikasikan benar, 4 sebagai kelas A, 1 diklasifikasikan sebagai kelas B dan 2 diklasifikasikan sebagai kelas C.
5	Nilai {4 2 7 0 14} menunjukkan ada (4+2+7+0+14) instances kelas C, 14 diantaranya diklasifikasikan benar, 4 sebagai kelas D, 2 diklasifikasikan sebagai kelas A dan 7 diklasifikasikan sebagai kelas B.

Tabel 4. *Evaluasi Model*

Kriteria	Nilai
Correctly Classified Instances	72
Incorrectly Classified Instances	28
Accuracy (%)	72%
Kappa statistic	0.6217
Mean absolute error	0.1261
Root mean squared error	0.2511
Relative absolute error	41.9858 %
Root relative squared error	64.8972 %

Tingkat akurasi model yang dihasilkan sebesar 72%. Jika dilihat dari nilai kappa, model yang dihasilkan memiliki reliabilitas sedang [8]. Reliabilitas sedang belum menunjukkan baik, hal ini disebabkan kondisi data tidak stabil dari subyek yang diukur.
Tabel 4 menunjukkan *true positive rate (TP Rate), false positive rate (FP Rate) , Precision , Recall , F-Measure* dan *ROC Area* pada masing-masing kelas.

Tabel 5.*Nilai Kinerja Per Kelas*

Kelas	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*ROC Area*
A	0,864	0,103	0,704	0,864	0,776	0,965
B	0,882	0,167	0,732	0,882	0,800	0,948
C	0,519	0,055	0,778	0,519	0,622	0,917
D	0,889	0,055	0,615	0,889	0,727	0,977
E	0,125	0,000	1,000	0,125	0,222	0,572

Sebuah sistem dianggap baik jika nilai *precision* dan *recall* nya tinggi (dinyatakan dalam bentuk nilai 1-100% atau 0-1). Sedangkan nilai ROC, memiliki tingkat nilai untuk mengukur kinerja sistem [9] :

a. Nilai 0,90 – 1,00 = *excellent classification*
b. Nilai 0,80 – 0,90 = *good classification*
c. Nilai 0,70 – 0,80 = *fair classification*
d. Nilai 0,60 – 0,70 = *poor classification*
e. Nilai 0,50 – 0,60 = *failure classification*

Dari nilai *precision*, *recall* dan ROC, dapat dilihat kinerja terburuk sistem adalah dalam melakukan klasifikasi pada kelas E, walaupun nilai precision nya sangat tinggi, namun pada recall dan ROC sangat buruk.
Dari evaluasi model, didapatkan 4 klasifikasi benar dan 4 klasifikasi salah dari 8 data untuk kelas E.
Kinerja sistem yang terburuk selanjutnya adalah dalam klasifikasi kelas C dengan 14 klasifikasi benar dan 13 klasifikasi salah dari 27 data.

3. Kesimpulan

Algoritma ID3 memberikan performansi yang cukup baik pada kasus klasifikasi performansi mahasiswa pada studi kasus ini. Dapat dilihat pada akurasi model sebesar 72%. Namun, dari nilai kinerja sistem dalam melakukan klasifikasi per kelas, masih buruk pada dua kelas. Hal ini

dapat disebabkan karena jumlah data latih yang sedikit serta kondisi data yang kurang seragam.
Pada penelitian selanjutnya, perlu dikaji lebih lanjut dengan kondisi data yang lebih banyak serta menyaring data yang kondisinya tidak stabil agar semakin meningkatkan reliabilitas model. Selain itu, dapat juga dilakukan penerapan metode klasifikasi selain algoritma ID3 untuk menemukan metode yang terbaik pada studi kasus ini.

Daftar Pustaka

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, no. Second Edition. 2006.
[2] Ogunde and Ajibade, "A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm" *Comput. Sci. Inf. Technol.*, vol. 2, no. 1, pp. 21–46, 2014.
[3] A. K. M. Pal and S. Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students," *Int. Journal. Electron. Commun. Comput. Eng.*, vol. 4, no. 5, pp. 1560–1565, 2013.
[4] B. K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification," *Int. Journal Comput. Sci. Inf. Secur.*, vol. 9, no. 4, 2011.
[5] Kristanto, Obbie, "Penerapan Algoritma Klasifikasi Data Mining ID3 Untuk Menentukan Penjurusan Siswa Sman 6,"Udinus Repository. 2013.
[6] Marsland, Stephen. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall. 2011
[7] Mayadewi, Pramita and Ely Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining" Seminar Nasional Sistem Informasi Indonesia (SESINDO). 2015
[8] B. Murti, "Validitas dan Reliabilitas Pengukuran," *Matrikulasi Program Studi. Doktoral*, pp. 1–19, 2011.
[9] Gorunescu, F. Data Mining Concept, Model and Techniques. Berlin : Springer. 2011

Biodata Penulis

Andika Elok Amalia, memperoleh gelar Sarjana Teknik (S.T), Jurusan Teknik Informatika Institut Teknologi Telkom (sekarang Telkom University) dan memperoleh gelar Magister Teknik (M.T) Magister Teknik Elektro Institut Teknologi Bandung. Saat ini menjadi Dosen di ST3 Telkom Purwokerto.

Muhammad Zidny Naf'an, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika UIN Syarif Hidayatullah dan memperoleh gelar Magister Komputer (M.Kom) Magister Ilmu Komputer Universitas Indonesia. Saat ini menjadi Dosen di ST3 Telkom Purwokerto.

ISSN : 2302-3805