PEMANFAATAN *EDUCATIONAL DATA MINING* (EDM) UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS: TI-S1 UDINUS)

Defri Kurniawan[1*], Wibowo Wicaksono[1] dan Yani Parti Astuti[1]
[1]Jurusan Teknik Informatika, Universitas Dian Nuswantoro
Jl. Imam Bonjol No 207 – Semarang, Indonesia
[*]*Email: defri.kurniawan@dsn.dinus.ac.id*

Abstrak

Tersedianya data yang melimpah pada institusi pendidikan harus dimanfaatkan dengan baik. Menemukan pola studi mahasiswa dan hubungan antar atribut-atribut data pendidikan yang mempengaruhi masa studi mahasiswa dalam suatu data besar, menjadi kajian dalam penelitian ini. Data mining dapat diusulkan sebagai salah satu pendekatan yang dapat dilakukan untuk memprediksi kinerja siswa. Algoritma C4.5 diterapkan untuk menemukan pola klasifikasi terhadap mahasiswa yang telah lulus tepat waktu dan tidak tepat waktu serta melakukan prediksi terhadap data uji yang diberikan. Hasil akurasi menunjukkan algoritma C4.5 mampu melakukan prediksi dengan baik (73,68%) terhadap masa studi mahasiswa yang tepat waktu dan tidak tepat waktu. Penerapan Data Mining pada bidang pendidikan (Educational Data Mining) memberikan kemajuan dan kontribusi besar pada dunia pendidikan dan pada bidang riset data mining.

***Kata kunci**: Data Mining, EDM, Masa Studi Mahasiswa, Algoritma C4.5, Decision Tree*

PENDAHULUAN

Salah satu cara untuk mencapai tingkat kualitas tertinggi dalam sistem pendidikan tinggi adalah dengan menemukan pengetahuan dari data pendidikan untuk mempelajari atribut utama yang dapat mempengaruhi kinerja siswa (Abu Tair & Al-Helees, 2012). Masa studi merupakan atribut penting bagi pengelola akademik, dengan dapat memprediksi masa studi mahasiswa, pihak universitas dapat meminimalisir kegagalan kelulusan mahasiswa dengan membuat perencanaan, pengawalan studi dan bimbingan lebih intensif. Mahasiswa yang memiliki masa studi lebih (tidak tepat waktu) memiliki potensi lebih besar gagal dibandingkan dengan mahasiswa yang dapat menyelesaikan studinya dengan tepat waktu.

Tersedianya data yang melimpah pada institusi pendidikan harus dimanfaatkan dengan baik. Namun sulitnya memahami dan menemukan hubungan atribut-atribut data yang mempengaruhi hasil masa studi mahasiswa yang dapat digolongkan sebagai tepat waktu dan tidak tepat waktu, menjadi kajian dalam penelitian ini.

Menganalisa kinerja mahasiswa (*student performance*), mengidentifikasi keunikan-keunikan yang ada pada mahasiswa dan membangun suatu strategi pengembangan lebih lanjut serta tindakan-tindakan yang dapat dilakukan untuk masa mendatang, merupakan tantangan utama bagi universitas modern saat ini (Kabakchieva, 2013). *Data mining* dapat diusulkan sebagai salah satu pendekatan yang dapat dilakukan untuk memprediksi kinerja siswa (Osmanbegovic & Suljic, 2012). Kinerja siswa dalam hal ini adalah capaian kelulusan studi mahasiswa dimana mereka dapat lulus tepat waktu atau tidak tepat waktu.

Data mining merupakan suatu cara dalam menggali informasi dari sejumlah data yang biasanya tersimpan dalam repositori dengan menggunakan teknologi pengenalan pola, statistik dan teknik matematika (Larose, 2006). Klasifikasi dan Prediksi merupakan pekerjaan-pekerjaan yang dapat dilakukan pada *data mining*. Klasifikasi adalah proses menemukan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep (Han & Kamber, 2007). Model tersebut akan digunakan untuk melakukan prediksi *output* terhadap sekumpulan data yang belum diketahui label kelasnya.

Penerapan metode *data mining* dalam menganalisis data yang tersedia di lembaga pendidikan didefinisikan sebagai *Educational Data Mining* (EDM) (Romero & Ventura, 2007). EDM merupakan suatu aliran yang relatif baru dalam penelitian *data mining*. EDM menggunakan beberapa teknik seperti *Decision Trees, Neural Networks, Naïve Bayes, K-Nearest Neighbor* dan lainnya (Yadav & Pal,

2012). EDM berkaitan dengan pengembangan metode untuk mengeksplorasi jenis yang unik dari data-data pada pengelolaan pendidikan dan menggunakannya untuk lebih memahami siswa dan pengelolaannya (Baker, 2010). Hal tersebut merupakan tujuan yang ingin dicapai dalam pemanfaatan *data mining* di bidang pendidikan.

Pada penelitian Pandey dan Pal (Pandey & Pal, 2011) EDM digunakan untuk mengukur kinerja siswa pendatang baru, apakah mereka bisa menjalankan studinya dengan baik (*performed*) atau tidak dengan memilih 600 mahasiswa dari perguruan tinggi yang berbeda dari Dr. R. M. L. Awadh University, Faizabad, India dengan menggunakan *Byes Classification*.

Bharadwaj dan Pal (B.K & S, 2011) melakukan penelitian pada kinerja siswa dengan memilih 300 mahasiswa dari 5 perguruan tinggi sederajat yang berbeda pada BCA (*Bachelor of Computer Application*) dari Dr. R. M. L. Awadh University, Faizabad, India dengan menggunakan metode klasifikasi *Bayesian* pada 17 atribut, ditemukan bahwa faktor-faktor seperti ujian SLTA, lokasi tinggal, media pengajaran, kualifikasi ibu, kebiasaan lain mahasiswa, pendapatan tahunan keluarga dan status keluarga siswa tersebut sangat terkait dengan prestasi akademik siswa.

Penelitian oleh Z. J. Kovacic (Z. J, 2010) berdasarkan studi kasus mengidentifikasi sampai sejauh mana data pendaftaran dapat digunakan untuk memprediksi keberhasilan siswa. Algoritma CHAID dan CART diterapkan pada data pendaftaran mahasiswa Sistem Informasi politeknik terbuka New Zealand untuk mendapatkan dua pohon keputusan dalam mengelompokkan siswa sukses dan tidak sukses. Akurasi diperoleh masing-masing untuk CHAID dan CART adalah 59,4 dan 60,5**.**

Penelitian Yadav dan Pal (Yadav & Pal, 2012) melakukan prediksi pada data pendidikan untuk mengidentifikasi siswa yang lemah dan membantu mereka untuk mencetak nilai yang lebih baik. Algoritma C4.5, ID3 dan CART diterapkan dan dibandingkan akurasinya, hasil menunjukkan bahwa teknik C4.5 memiliki akurasi paling tinggi yaitu 67,78% dibandingkan dengan teknik lainnya.

Berdasarkan penelitian-penelitian yang telah dilakukan sebelumnya, algoritma C4.5 akan digunakan pada penelitian ini karena memiliki tingkat akurasi yang lebih baik dan dapat memberikan gambaran klasifikasi mahasiswa yang tepat waktu atau tidak tepat waktu berupa pohon keputusan (*Decision Tree*) yang bermanfaat bagi pengelola akademik

Decision Tree menyerupai sebuah struktur pohon dimana terdapat *node* internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas (Han & Kamber, 2007). Pohon keputusan bekerja mulai dari akar paling atas (*root node*), jika diberikan sejumlah data uji, misalnya X dimana kelas dari data X belum diketahui, maka pohon keputusan akan menelusuri mulai dari akar sampai *node* dan setiap nilai dari atribut sesuai data X diuji apakah sesuai dengan aturan *Decision Tree*, kemudian pohon keputusan akan memprediksi kelas dari tupel X.

METODOLOGI

Metode penelitian yang dilakukan adalah metode penelitian eksperimen dengan tahapan penelitian sebagai berikut (Santoso, 2007):

Tahap Pengumpulan Data

Data set yang digunakan dalam penelitian ini adalah data mahasiswa Teknik Informatika (TI) S-1 angkatan 2008, 2009, 2010, 2011 sejumlah 1473 *records*. Data mahasiswa yang diambil adalah data identitas mahasiswa yang menggambarkan informasi profil mahasiswa dan data akademik yang menggambarkan informasi akademik mahasiswa berupa IPK dan Masa Studi

Pengolahan Awal Data

Pengolahan awal data diperlukan untuk proses penyederhanaan data, agar data tersebut dapat dikenali dan digunakan dalam algoritma yang diusulkan. Proses pengolahan awal data tersebut adalah:

1. *Data Integration* yaitu menyatukan tempat penyimpanan. Data identitas dan mahasiswa yang diperoleh disatukan dalam satu media penyimpanan.
2. *Data reduction* yaitu untuk memperoleh data yang mempunyai atribut dan *record* yang lebih sedikit dengan cara mengurangi

record yang tidak diperlukan atau yang tidak terisi.

Pada *data reduction*, data yang tidak terisi selanjutnya dieliminasi yaitu dari atribut lokasi tinggal dan nama sekolah asal yang sering sekali tidak terisi. Atribut lokasi tinggal digunakan untuk menentukan status tinggal mahasiswa bersama orang tua atau tidak bersama orang tua. Atribut sekolah asal digunakan untuk mengkategorikan jenis sekolah SMA, SMK, Sekolah Lanjut, MA, Pesantren. Hasil pengolahan data awal (*preprocessing*) mengahasilkan data valid sejumlah 948 *records* serta atribut-atribut yang digunakan dalam penelitian yang tersaji pada tabel 1

Tabel 1. Atribut-atribut Data Pada Penelitian

Atribut	Deskripsi	Nilai
Jenis Kelamin	Menjelaskan jenis kelamin mahasiswa laki-laki (L) atau perempuan (P).	L atau P
Jenis Sekolah Asal	Menjelaskan jenis sekolah asal dari mahasiswa yang bersangkutan.	SMA, SMK, MA, Pesantren, Sekolah Lanjut
Status Tinggal	Menjelaskan status tinggal mahasiswa. Apakah tinggal bersama orang tuanya atau tidak.	Bersama Orang Tua atau Tidak Bersama Orang Tua
Pekerjaan Orang Tua Wali (*Job*)	Menjelaskan status pekerjaan orang tua wali mahasiswa.	PNS, Swasta, TNI/POLRI, Wirausaha, Petani/ Peternak, Lainnya
IPK	Menjelaskan nilai Index Prestasi Komulatif (IPK)	0 s.d 4,00
Status Masa Studi	Atribut Status Masa Studi merupakan variabel *output* atau label. Jika masa studi lebih besar dari empat tahun maka Tidak Tepat Waktu. Namun Jika kurang dari sama dengan empat, maka Tepat Waktu	Tepat Waktu atau Tidak Tepat Waktu

Model/Metode Yang Diusulkan

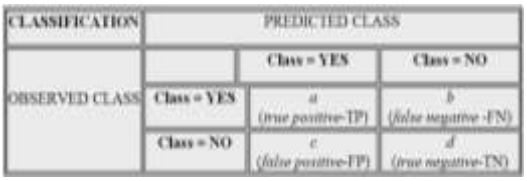
Model/metode yang diusulkan dalam penelitian ini menggunakan pembelajaran pohon keputusan (*Decision Tree Learner*) dengan Algoritma C4.5.

Algoritma C4.5 merupakan penerus dari ID3 yang dikembangkan oleh Quinlan Ross (J. R, 1992). Langkah awal algoritma C4.5 adalah dengan menghitung nilai *gain ratio* dari setiap atribut. Nilai *gain ratio* tertinggi akan menjadi simpul akar (*root node*). C4.5 akan menghilangkan cabang yang tidak perlu dalam pohon keputusan untuk meningkatkan akurasi klasifikasi (Yadav & Pal, 2012). Algoritma C4.5, ID3 dan CART termasuk dalam pembelajaran pohon keputusan (*Decision Tree Learner*).

Eksperimen Dan Pengujian Model

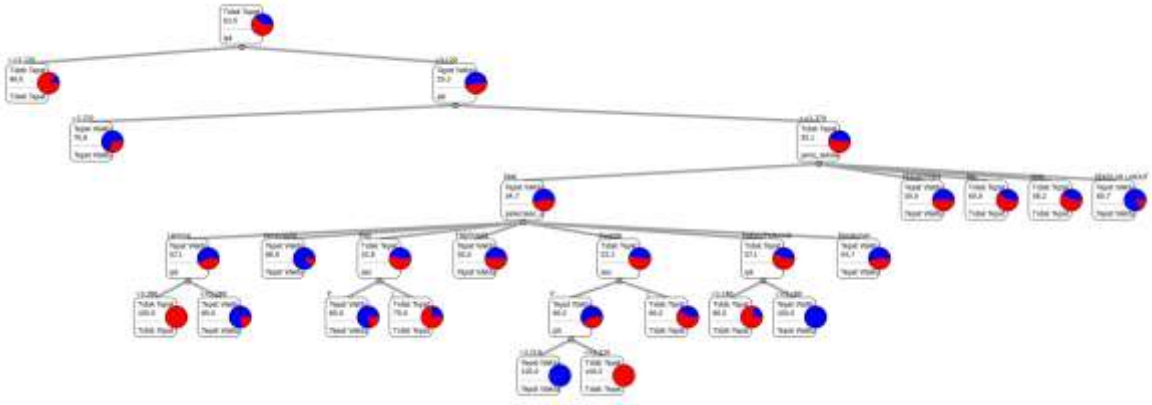
Tahapan eksperimen dan pengujian model pada penelitian ini adalah:

1. Menyiapkan data untuk melakukan eksperimen.
2. Pengolahan awal data (*preprocessing)* dengan mereduksi data – data yang kosong.
3. Implementasi *data mining* menggunakan bantuan *software Orange* versi 2.0b untuk membangun model klasifikasi algoritma C4.5. *Orange* merupakan *free software* dengan model perangkat lunak berbasis komponen untuk *machine learning* dan *data mining* yang dikembangkan pada *Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana*, Slovenia, bersama dengan komunitas *open source* http://orange.biolab.si/
4. Menguji model algoritma C4.5 dengan menghitung nilai akurasi klasifikasi dengan *confusion matrix.*



Gambar 1. *Confusion Matrix* Kasus Dua Kelas Model

Kolom a (*true positive*-TP) dan d (*true negative*-TN) merupakan klasifikasi yang benar, dimana *classifier* memprediksi secara tepat dengan kondisi sebenarnya. Sedangkan Suatu *false negative*-FN / kolom b adalah suatu kondisi yang salah prediksi, ketika diperkirakan sebagai no (*negative*) namun hasil sebenarnya *yes* atau positif. Sedangkan *false positive*-FP / kolom c adalah suatu kondisi salah yaitu ketika diperkirakan *yes* atau positif, namun sebenarnya *no* atau *negative* (Han & Kamber, 2007). Berdasarkan empat kondisi



Gambar 4. Hasil Pohon Keputusan (*Decision Tree*) Kelas Tepat Waktu dan Tidak Tepat Waktu

yang dihasilkan *confusion matrix*, nilai akurasi klasifikasi dapat dihitung sesuai dengan rumus (1).

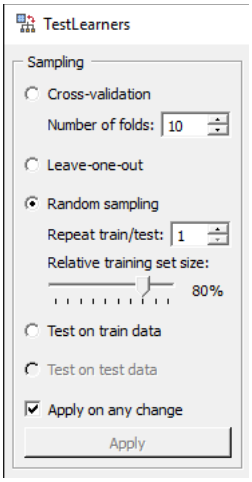
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

5. Menganalisa hasil dari penggunaan algoritma C4.5

HASIL DAN PEMBAHASAN

Pada implementasi *data mining*, data dibagi menjadi 2 (dua) yaitu data *training* dan data *testing*. Data *training* digunakan sebagai data pelatihan untuk membangun model klasifikasi berdasarkan algoritma C4.5. Data *testing* digunakan sebagai pengujian untuk mengevaluasi kinerja dari algoritma yang digunakan.

Pada penelitian ini menggunakan *random sampling* untuk memilih data secara acak yang digunakan sebagai data *training* dan data *testing* dengan pembagian data *training* sebesar 80% dari 948 data yaitu 758 data. Data *testing* sebesar 20% dari 948 data yaitu 190 data.



Gambar 2. Penerapan *Random Sampling* Dengan *Orange Software*

Data *testing* selanjutnya diuji dengan model klasifikasi yang telah dibangun dari data *training* untuk memprediksikan tingkat akurasi dari data pengujian yang digunakan. Akurasi klasifikasi didapatkan berdasarkan tabel *confusion matrix*. *Confusion matrix* dari data testing yang digunakan dengan keluaran Tepat Waktu dan Tidak Tepat Waktu sepeti pada gambar dibawah ini:

	Prediction		
	Tepat Waktu	Tidak Tepat Waktu	
Tepat Waktu	42	28	**70**
Tidak Tepat Waktu	22	98	**120**
	64	**126**	**190**

Gambar 3 Hasil Tabel *Confusion Matrix* Menggunakan *Orange Software*

Berdasarkan tabel *confusion matrix* didapatkan tingkat akurasi klasifikas algoritma C4.5 sebesar 73,68%.

Tabel 2. Hasil Akurasi Algoritma C4.5

Algoritma	Akurasi
C4.5	73,68%

Algoritma C4.5 menghasilkan bentuk pohon keputusan (*decision tree*) seperti yang ditunjukkan gambar 4. Terlihat bahwa IPK merupakan atribut paling menentukan (*root node*) dari atribut-atribut lainnya. Warna merah mewakili Kelas Tidak Tepat Waktu dan warna biru mewakili Kelas Tepat Waktu.

KESIMPULAN

Pada penelitian ini, penggunaan algoritma C4.5 mampu melakukan prediksi dengan baik (73,68%) terhadap masa studi mahasiswa yang tepat waktu dan tidak tepat waktu. Pembentukan pohon keputusan (*Decision Tree*) dapat digunakan oleh pengelola akademik di dalam memetakan mahasiswa yang berpotensi mengalami keterlambatan masa studi di masa mendatang. Penerapan *Educational Data Mining* (EDM) memberikan kemajuan dan kontribusi pada dunia pendidikan dan pada bidang riset *data mining*.

DAFTAR PUSTAKA

Abu Tair, M. M., & Al-Helees, A. M. 2012, February, Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information and Communication Technology Research, 2*.

B.K, B., & S, P., 2011, Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security (IJCSIS), 9*(4), pp. 136-140.

Baker, R., 2010, *Data Mining for Education* (3rd Edition ed.). UK: Elsevier.

Han, J., & Kamber, M., 2007, *Data Mining Concepts and Techniques* (2nd ed.). San Francisco, United State America: Morgan Kaufmann Publishers.

J. R, Q., 1992, *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc.

Kabakchieva, D., 2013, Predicting Student Performance by Using Data Mining. *Cybernetics and Information Technologies*.

Larose, D. T., 2006, *Data Mining Methods and Models.* Hoboken, New Jersey, United State of America: John Wiley & Sons, Inc.

Osmanbegovic, E., & Suljic, M., 2012, May, Data Mining Approach For Predicting Student Performance. *Journal of Economics and Business, X*(1).

Pandey, U., & Pal, S., 2011, Data Mining: A prediction of performer or underperformer using classification. *(IJCSIT) International Journal of Computer Science and Information Technology, 2(2)*(ISSN:0975-9646), 686-690.

Romero, C., & Ventura, S., 2007, Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transactions On Systems, Man, And Cybernatics*.

Santoso, B., 2007, *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Yadav, S. K., & Pal, S., 2012, Data Mining A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology Journal (WCSIT), 2*, 51-56.

Z. J, K., 2010, Early prediction of student success: Mining student enrollment data. *Proceedings of Informing Science & IT Education Conference.*