PREDIKSI PRESTASI AKADEMIK MAHASISWA MENGGUNAKAN ALGORITMA *RANDOM FOREST* DAN C4.5

Safitri Linawati[1]; Siti Nurdiani[2];Kartika Handayani[3]; Latifah
Ilmu Komputer [1][2];Teknik Komputer[3];SistemInformasi Akuntansi [4]
STMIK Nusa Mandiri[1][2]; Universitas Bina Sarana Informatika[3] [4]
Nusamandiri@ac.id; www.bsi.ac.id
safitrilw24@gmail.com[1];siti.sxd@musamandiri.ac.id[2];kartika.kth@bsiac.id[3]; Latifah.lat@bsi.ac.id[4]

Abstract— *In the first academic year the average student learning outcomes cause various problems that will have an impact on the following semester. Evaluation of the successful implementation of the study program at a tertiary institution is seen from the results of student learning. Data mining methods might be able to identify the right model to fix every problem that arises so that it can be handled by the lecturer concerned. The purpose of this paper is to apply data mining methods to predict academic achievement of students by exploring various parameters. We propose efficient data mining techniques to predict student academic achievement. With classification techniques from data mining, namely Decision Tree C4.5 and Random Forest. The research phase is to do pre-processing on the dataset that is coding the non-numeric attribute values, then cross-validation testing is performed. And to build a predictive model of student academic achievement, we carry out a classification process using the Decision Tree C4.5 and Random Forest methods. The testing method used in this study is Confusion Matrix, a method that is usually used in calculating the accuracy of data mining. The resulting calculation consists of recall, accuracy and precision. The results of testing with the Decision Tree C4.5 and Random Forest classification models in this study indicate that the Random Forest classification model has an accuracy value of 92.4%, a precision of 91.4% and a recall of 92.4% higher than the Decision Tree C4.5.*

Keywords: *Student academic pretation, data mining, decision tree c4.5, random forest*

Intisari— Pada tahun akademik pertama rata-rata hasil belajar mahasiswa menimbulkan berbagai masalah yang akan berdampak pada semester selanjutnya. Evaluasi keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi dilihat dari hasil belajar mahasiswa. Metode data mining mungkin bisa mengidentifikasi model yang tepat untuk memperbaiki setiap permasalahan yang timbul agar dapat ditangani oleh dosen yang bersangkutan. Tujuan penelitian ini untuk menerapkan metode *data mining* untuk prediksi prestasi akademik mahasiswa dengan mengeksplorasi berbagai parameter. Kami mengusulkan teknik data mining yang efisien untuk memprediksi prestasi akademik mahasiswa. Dengan teknik klasifikasi dari data mining yaitu *Decision Tree C4.5* dan *Random Forest.* Tahap penelitian yang dilakukan adalah melakukan pre-processing pada dataset yaitu melakukan pemberian kode pada nilai atribut yang non-numerik, kemudian dilakukan pengujian *cross-validation*. Dan untuk membangun model prediksi prestasi akademik mahasiswa, kami melakukan proses klasifikasi menggunakan metode Decision Tree C4.5 dan Random Forest. Metode yang digunakan dalam pengujian penelitian ini yaiut *Confusion Matrix* yaitu suatu metode yang biasanya digunakan dalam melakukan perhitungan akurasi pada suatu data mining. Perhitungan yang dihasilkan terdiri *recall, acuraccy* dan precision. Hasil dari pengujian dengan model klasifikasi *Decision Tree C4.5* dan *Random Forest* pada penelitian ini menunjukkan bahwa model klasifikasi Random Forest memiliki nilai *accuracy* sebesar 92.4%, *precision* sebesar 91.4% dan *recall* sebesar 92,4% lebih tinggi dibandingkan *Decision Tree C4.5.*

Kata Kunci: *Pretasi akademik mahasiswa, data mining, decision tree c4.5, random forest*

PENDAHULUAN

Pada tahun akademik pertama rata-rata hasil belajar mahasiswa menimbulkan berbagai masalah yang akan berdampak pada semester selanjutnya. Evaluasi keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi dilihat dari hasil belajar mahasiswa. Ukuran yang kami gunakan dalam menghitung prestasi akademik mahasiswa yaitu dengan melihat Indeks Prestasi Mahasiswa (IPK).

Sebuah lembaga pendidikan tinggi memberikan konsep pembelajaran yang baik kepada mahasiswanya agar dapat mencapai

prestasi akademik yang memuaskan. Pada kenyataannya prestasi akademik mahasiswa sangat bervariasi, bahkan ketika mereka belajar dalam metode pendidikan yang sama dan diberikan lingkungan dan alat yang sama.

Metode data mining mungkin bisa mengidentifikasi model yang tepat untuk memperbaiki setiap permasalahan yang timbul agar dapat ditangani oleh dosen yang bersangkutan. Data mining adalah sebuah kajian dari pengamatan untuk data dalam jumlah yang besar untuk mendapatkan suatu hubungan yang tidak diketahui sebelumnya dan ada dua metode baru untuk mempersingkat data agar mudah dipahami serta berguna untuk pemilih data (Jefri & Kusrini, 2013). Selain menganalisa obesrvasi data dalam jumlah besar, data mining juga dapat menangani dimensi data yang tinggi dan juga data yang memiliki sifat berbeda.

Beberapa peneliti telah membuat kontribusi signifikan untuk bidang ini. Jai dan K (2015) membangun model prediksi menggunakan algoritma Multi-layer perceptron dengan atribut akademik, pribadi dan ekonomi. Akurasi rata-rata dicapai dengan semua atribut adalah 52% dan dengan atribut yang dipilih adalah 33% (Ruby & David, 2015).

Mustafa, Ramadhan dan Thenata (2018) melakukan evaluasi kinerja akademik mahasiswa STMIK Dipanegara Makassar untuk membuat bentuk sebuah tabel kemungkin sebagai proses dasar pengelompokkan prestasi pembelajaran mahasiswa yang kelulusannya akan dikelompokkan dan memberikan pengajuan untuk proses kelulusan tepat waktu yang dengan nilai optimal berdasarkan hasil nilai yang telah ditempuh mahasiswa. Pengujian pada beberapa data mahasiswa angkatan 2008- 2011 yang diambil secara acak, algoritma NBC menghasilkan nilai akurasi 92,3%.

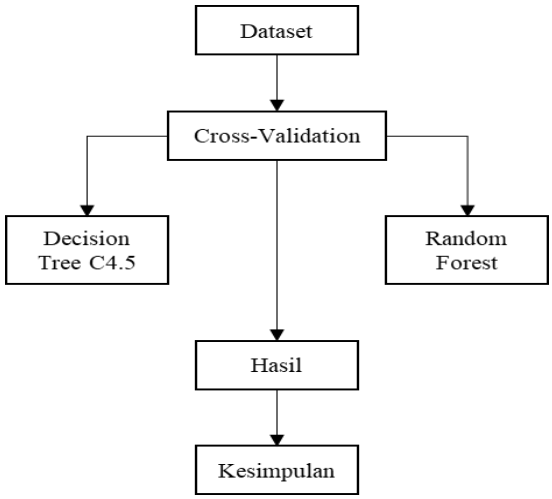
Tujuan penelitian ini untuk menerapkan algoritma Decision Tree C4.5 dan Random Forest untuk prediksi prestasi akademik mahasiswa dengan mengeksplorasi berbagai parameter yaitu kategorisasi angka numerik, jumlah kategori Indeks Prestasi Mahasiswa.

Dalam penelitian ini, kami mengusulkan teknik data mining yang efisien untuk memprediksi prestasi akademik mahasiswa. Dengan teknik klasifikasi dari data mining yaitu Decision Tree C4.5 dan Random Forest kemudian, pengklasifikasian diusulkan dengan *dataset* untuk melakukan proses data latih dan model yang diperoleh. Model yang diperoleh disertakan dengan data uji. *Dataset* dikumpulkan dari Database Bagian Akademik Institut Bisnis & Multimedia asmi. Hasil yang diinginkan adalah model dari teknik klasifikasi terbaik dari 2 algoritma tersebut dan membandingkan nilai akurasi.

METODE PENELITIAN

Sumber data untuk membangun teknik yang diusulkan dalam memprediksi prestasi akademik mahasiswa terdiri dari 170 record dan 11 fitur yang berbeda. Dengan mengeksplorasi berbagai parameter, yaitu kategorisasi angka numerik dan jumlah kategori Indeks Prestasi Mahasiswa.

Dalam penelitian ini, kami menggunakan aplikasi Weka yang mendasari pendekatan-pendekatan *machine learning*. *Dataset* tersebut nantinya akan dilakukan *test options 10 fold cross-validation.* Kemudian kami melakukan proses klasifikasi menggunakan metode Decision Tree C4.5 dan Random Forest pada *dataset* untuk membangun model prediksi prestasi akademik mahasiswa dan menghasilkan nilai akurasi yang tinggi. Metode yang diusulkan pada penelitian ini memiliki beberapa tahap.



Gambar 1. Tahap Penelitian

HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan mengenai tahapan penelitian untuk memperoleh hasil klasifikasi prediksi prestasi akademik mahasiswa menggunakan metode Decision Tree C4.5 dan Random Forest. adapun tahapan tersebut adalah tahap persiapan *dataset*, pengujian *10 fold cross-validation* dan klasifikasi. *Dataset* yang digunakan adalah data mahasiswa angkatan tahun akademik 2018/2019 yang diambil dari Bagian Akademik Institu Bisnis dan Multimedia asmi.

1. Persiapan *dataset*

   Pra-processing merupakan beberapa langkah penting dalam proses pengambilan data seperti memberikan sebuah jarak variabel dan beberapa jenis pengkodean. Sebagai

contoh, satu fitur dengan range [0, 1] dan yang lain dengan nilai [-100, 100] tidak akan memiliki beban yang sama pada teknik yang digunakam dan akan berpengaruh pada hasil akhir *data mining*. Oleh karena itu, diajukan untuk pemberian jarak dan membawa karakteristik tersebut ke beban yang sama untuk ditinjau lebih lanjut (Kantardzic, 2011).

Pada dataset beberapa atribut memiliki nilai tipe non-numerik, jadi dalam penelitian ini kami pertama-tama mengganti nilai tipe non-numerik menjadi nilai numerik. Dalam proses mengubah nilai non-numerik menjadi nilai numerik menggunakan teknik pengkodean untuk mendekripsi data dengan aturan yang layak dan mengusulkan model data mining untuk membuat prediksi berdasarkan data yang telah didekripsi. Atribut yang memiliki nilai non-numerik ditunjukkan pada tabel dibawah ini:

Tabel 1. Atribut dengan nilai non-numerik

Atribut	**Deskripsi**	**Nilai**
Jenjang	Jenjang pendidikan perguruan tinggi	A=S3, B=S2, C=S1, D=D4, E=D3, F=D1
Prodi	Program studi perguruan tinggi	Sekretari, Administrasi Bisnis
Asal	Asal tempat lahir	Jakarta, Luar Jakarta
JK	Jenis Kelamin	Laki-laki (L), Perempuan (P)

Terdapat 4 atribut yang memiliki nilai non-numerik. Selanjutnya, cukup menetapkan nilai non-numerik menjadi numerik. Pada proses ini, kami tidak melakukan analisis yang lebih dalam pada tahap pertama, karena hanya perlu melihat apakah akan ada perbedaan antara data yang dienkripsi dan data yang didekripsi. Dengan kata lain, pada tahap ini bertujuan untuk mengenkripsi data dengan mengganti data mentah dengan angka (Zhang, 2018).

Tabel 2. Pengkodean nilai atribut Jenjang menjadi nilai non-numerik

Atribut	Nilai Non-numerik	Nilai Numerik
Prodi	Diploma I	1
	Diploma III	2
	Diploma IV	3
	Strata 1	4
	Strata 2	5
	Strata 3	6

Tabel 3. Pengkodean nilai atribut Prodi menjadi nilai non-numerik

Atribut	Nilai Non-numerik	Nilai Numerik
Prodi	Administrasi Bisnis	1
	Akuntansi	2
	Komunikasi	3
	Manajemen	4
	Sekretari	5
	Sistem Informasi	6

Tabel 4. Pengkodean nilai atribut Asal menjadi nilai non-numerik

Atribut	Nilai Non-numerik	Nilai Numerik
Asal	Jakarta	1
	Luar Jakarta	2

Tabel 5. Pengkodean nilai atribut JK menjadi nilai non-numerik

Atribut	Nilai Non-numerik	Nilai Numerik
JK	Laki-laki	1
	Perempuan	2

Pada atribut Jenjang memiliki nilai perpaduan antara non-numerik dan numerik, nilai-nilai tersebut independen dan tidak saling mempengaruhi, jadi kami menetapkan dengan angka 1 – 6 dimulai dari jenjang terendah sampai tertinggi. Pada atribut job memiliki nilai-nilai independen yang tidak saling mempengaruhi, kami menetapkan dengan angka yang berbeda dari 1 – 6. Pada atribut Asal, memiliki jenis yang independen dan tidak saling mempengaruhi, jadi kami menggunakan angka numerik yang berbeda, Nilai 1 untuk asal Jakarta dan nilai 2 untuk luar Jakarta. Pada atribut JK, juga memiliki jenis yang independen dan tidak saling mempengaruhi, jadi kami menggunakan angka numerik yang berbeda, Nilai 1 untuk laki-laki dan nilai 2 untuk perempuan.

Selanjutnya data disiapkan dengan melakukan seleksi dan transformasi *dataset*. Pemilihan data bertujuan untuk mengumpulkan variabel-variabel relevan yang digunakan dalam penelitian ini. Dan perubahan data digunakan untuk mengubah *dataset* sehingga konten informasi terbaik diambil dan dimasukkan pada Weka yang digunakan dalam format yang tepat.

Tahap berikutnya menentukan atribut yang dijadikan sebagai pola untuk memprediksi prestasi akademik mahasiswa.

Atribut yang digunakan dalam penelitian ini diambil berdasarkan referensi dari penelitian sebelumnya.

Tabel 6. Transformasi *Dataset*

Atribut	Deskripsi	Nilai
Jenjang	Jenjang pendidikan perguruan tinggi	6=S3, 5=S2, 4=S1, 3=D4, 2=D3, 1=D1
Prodi	Program studi perguruan tinggi	Administrasi Bisnis = 1, Akuntansi = 2, Komunikas = 3, Manajemen = 4, Sekretari = 5, Sistem Informasi = 6
Asal	Asal tempat lahir	Jakarta = 1, Luar Jakarta = 2
Umur	Umur saat masuk kuliah	<= 18 kategori 1, 19-21 kategori 2, >= 22 kategori 3
JK	Jenis Kelamin	Laki-laki = 1, Perempuan = 2
SKS1	SKS yang diambil 2 semester sebelumnya	<=12 sks kategori 0, 13-15 sks kategori 1, 16-18 sks kategori 2, 19-21 kategori 3, 22-24 kategori 4
SKS2	SKS yang diambil semester sebelumnya	<=12 sks kategori 0, 13-15 sks kategori 1, 16-18 sks kategori 2, 19-21 kategori 3, 22-24 kategori 4
IP1	Indeks Prestasi 2 semester sebelumnya	Ada 3 kategori IPK 1 = 0 – 2,50
IP2	Indeks Prestasi semester sebelumnya	1 = 2,51 – 3,50 2 = 3,51 – 4,00
IPK	Indeks Prestasi Kumulatif di semester berjalan	
IPK_Pred	Prediksi Indeks Prestasi Kumulatif semester selanjutnya	IPK-Tinggi IPK-Sedang-IPK-Rendah

2. Pengujian akurasi menggunakan *cross-validation*

Cross-validation merupakan suatu metode percobaan untuk penelaahan yang bepengalaman. Dimana dalam setiap lipatan dibagi menjadi beberapa bagian yaitu sebanyak 10 bagian dengan ukuran tiap bagian sama banyaknya. Setelah itu maka akan dilakukan data latih sebanyak 10 kali dengan menggunakan Sembilan lipatan untuk set data latih dan satu lipatan digunakan sebagai data uji. Kemudian dari setiap iterasi rata-rata kesalahan dari keseluruhan iterasi (Defiyanti Sofi, 2013).

3. Proses Klasifikasi

Pengelompokkan digunakan untuk menempatkan beberapa bagian yang tidak diketahui pada data dalam kelompok yang sudah diketahui. Pengelompokkan menggunakan variabel target dengan nilai nominal. Dalam satu set pelatihan, variabel target sudah diketahui. Dengan pembelajaran dapat ditemukan hubungan antara fitur dengan variabel target (Han & Kamber, 2006).

Pada penelitian ini kami mengusulkan model klasifikasi algoritma *Decision Tree C4.5* dan *Random Forest*. Setelah kami melakukan pre-processing pada dataset dan pengujian *cross-validation*, kemudian kami melakukan perbandingan dari model klasifikasi *Decision Tree C4.5* dan *Random Forest*. Percobaan dari penelitian dievaluasi dengan pengukuran *accuracy*, *precision* dan *recall*. Pengukuran yang dilakukan dengan menggunakan tabel pengelompokkan yang bersifat prediktif, disebut juga dengan *Confusion Matrix.*

Tabel 8. *Confusion Matrix*

True Values			
True	False		
TP	FP	True	Prediction
FN	TN	False	

a. Nilai *recall* menunjukkan berapa persen data kategori positif yang terklasifikasikan dengan benar oleh sistem. Nilai *recall* diperoleh dengan persamaan berikut:

$$Recall = \frac{TP}{FN+TP} * 100\% \text{ ..... (1)}$$

b. Nilai *accuracy* menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar. Dengan kata lain, nilai *accuracy* merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai *accuracy* dapat diperoleh dengan Persamaan berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \text{ .....(2)}$$

c. Nilai *precision* menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. *precision* dapat diperoleh dengan persamaan berikut:

$$Precision = \frac{TP}{FP+TP} * 100\% \text{ .....(3)}$$

dimana:

- TP *(True Positive)*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
- TN *(True Negative)*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN *(False Negative)*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP *(False Positive)*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem

4. Hasil Pengujian

Hasil pengujian model klasifikasi *Decision Tree C4.5* dan *Random Forest* dengan proses 10 *Fold Cross-validation*.

Tabel 9. *Confusion Matrix Decision Tree C4.5*

== *Confusion Matrix* ==			
IPK Rendah	IPK Menengah	IPK Tinggi	← *classified as*
1	9	0	IPK Rendah
5	119	3	IPK Menengah
0	5	28	IPK Tinggi

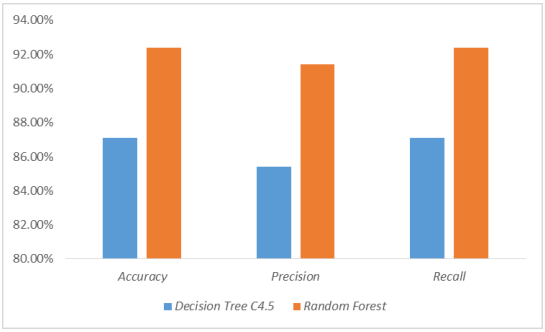
Tabel 10. *Confusion Matrix Random Forest*

== *Confusion Matrix* ==			
IPK Rendah	IPK Menengah	IPK Tinggi	← *classified as*
3	7	0	IPK Rendah
2	124	1	IPK Menengah
0	3	30	IPK Tinggi

Tabel 11. Hasil Pengujian *Accuracy, Precision, Recall*

Decision Tree C4.5 dan Random Forest

Classification	*Accuracy*	*Precision*	*Recall*
Decision Tree C4.5	87.1%	85.4%	87.1%
Random Forest	92.4%	91.4%	92.4%



Gambar 2. Grafik Hasil Pengujian *Accuracy, Precision, Recall*

Decision Tree C4.5 dan Random Forest

Hasil dari pengujian dengan model klasifikasi *Decision Tree C4.5* dan *Random Forest* pada penelitian ini menunjukkan bahwa model klasifikasi Random Forest memiliki nilai *accuracy, precision* dan *recall* lebih tinggi dibandingkan dengan *Decision Tree C4.5*.

1. KESIMPULAN

Berdasarkan hasil dan analisis pengujian yang telah dilakukan, maka dapat disimpulkan bahwa klasifikasi prediksi prestasi akademik mahasiswa menggunakan metode *Decisison Tree C4.5* menghasilkan akurasi 87.1 %, nilai presisi sebesar 85.4% dan nilai recall sebesar 87.1% , Sedangkan metode *Random Forest* menghasilkan akurasi ebesar 92.4%, nilai presisi sebesar 91.4% dan nilai recal sebesar 92.4%. Dengan kata lain metode *Random Forest* memiliki akurasi, presisi dan recall yang lebih baik dibandingkan dengan metode *Decision Tree C4.5*. Metode *Random Forest dapat* dapat digunakan untuk memprediksi Prestasi akademi mahasiswa.

REFERENSI

Defiyanti Sofi, M. K. (2013). Analisis dan Prediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining. *Syntak*.

Han, J., & Kamber, M. (2006). Data Mining Concept and Tehniques. In *San Fransisco: Morgan Kauffman.*

Jefri, & Kusrini. (2013). Implementasi Algoritma C4.5 Dalam Aplikasi Untuk Memprediksi Jumlah Mahasiswa Yang Mengulang Mata Kuliah Di STMIK AMIKOM Yogyakarta. *Naskah Publikasi*.

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. In *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*. https://doi.org/10.1002/9781118029145

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Creative Information Technology Journal*. https://doi.org/10.24076/citec.2017v4i2.106

Ruby, J., & David, K. (2015). Analysis of Influencing Factors in Predicting Students Performance Using MLP -A Comparative Study. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*. https://doi.org/10.15680/ijircce.2015.0302070

Zhang, J. (2018). Analysis of Neural Network on Bank Marketing Data.