PENERAPAN TEKNIK KLASIFIKASI UNTUK PREDIKSI KELULUSAN MAHASISWA BERDASARKAN NILAI AKADEMIK

I Made Budi Adnyana
Program Studi Sistem Informasi, ITB STIKOM Bali
budi.adnyana@stikom-bali.ac.id

ABSTRACT

Student graduation on time is an important aspect in assessing the quality of college. Prediction of graduation time is developed to support study program department in guiding their students to graduate on time. In this paper trying to apply classification techniques in data mining to solve this prediction problem. The classification methods used in this paper is a comparative approach between Naïve Bayes, J48, and Random Forest. Data sample obtained from SINAK system of STIKOM Bali, consist of 1580 graduation data and 41 courses that converted to attributes. The experiments are conducted using Weka application with 10-fold cross-validation. Experiment results show that Random Forest algorithm produce best accuracy (77.99%) and Naive Bayes algorithm has the worst accuracy (69.96%)
Keywords*: data mining, prediction, classification, graduation*

ABSTRAK

Kelulusan tepat waktu merupakan salah satu aspek penting dalam penilaian kualitas perguruan tinggi saat ini. Prediksi lama studi mahasiswa dirancang untuk mendukung prodi dalam membimbing mahasiswa agar lulus tepat pada waktunya. Pada penelitian ini mencoba menerapkan teknik klasifikasi pada data mining untuk memecahkan permasalahan prediksi lama studi tersebut. Metode klasifikasi yang digunakan dalam penelitian ini adalah komparasi antara Naïve Bayes, J48, dan Random Forest. Data sampel diperoleh dari sistem SINAK STIKOM Bali, terdiri dari 1580 data lulusan dan 41 matakuliah yang digunakan sebagai atribut. Uji coba dilakukan dengan menggunakan aplikasi WEKA dengan 10 folds cross-validation. Berdasarkan hasil uji coba menunjukkan algoritma Random Forest menghasilkan akurasi terbesar yaitu 77.99% dan paling rendah diperoleh dengan algoritma Naive Bayes yaitu 69.96%.
Kata kunci: data mining, prediksi, klasifikasi kelulusan

PENDAHULUAN

Perguruan tinggi saat ini dituntut untuk memiliki keunggulam bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Presentase mahasiswa tepat waktu dalam menyelesaikan studinya merupakan salah satu tolok ukur kualitas perguruan tinggi. Sehingga saat ini sangat menarik untuk diteliti tentang permasalahan kegagalan studi serta berbagai faktor penyebabnya (1). Rekam jejak nilai yang telah ditempuh oleh mahasiswa merupakan salah satu faktor yang berpengaruh terhadap lama studi mahasiswa. Perguruan tinggi perlu melakukan analisa terhadap kurikulum yang dijalankan untuk mengetahui matakuliah-matakuliah yang berkorelasi terhadap tingkat lama studi mahasiswa.

STIKOM Bali merupakan perguruan tinggi berbasis TI yang telah mengimplementasikan Sistem Informasi Akademik untuk menunjang berbagai proses akademis. Namun pemanfaatan data, khususnya data kelulusan dan data nilai mahasiswa, belum dilakukan secara optimal oleh pihak perguruan tinggi. Jika data ini dianalisa dan digali lebih dalam, maka bisa didapatkan suatu pola untuk melakukan prediksi terhadap lama studi mahasiswa. Prediksi lama studi ini salah satunya dapat diselesaikan dengan teknik klasifikasi pada data mining. Data mining merupakan serangkaian proses untuk mendapatkan suatu pola atau pengetahuan dari sekumpulan data. Data mining dalam bidang pendidikan dikenal dengan istilah Educational Data Mining (2).

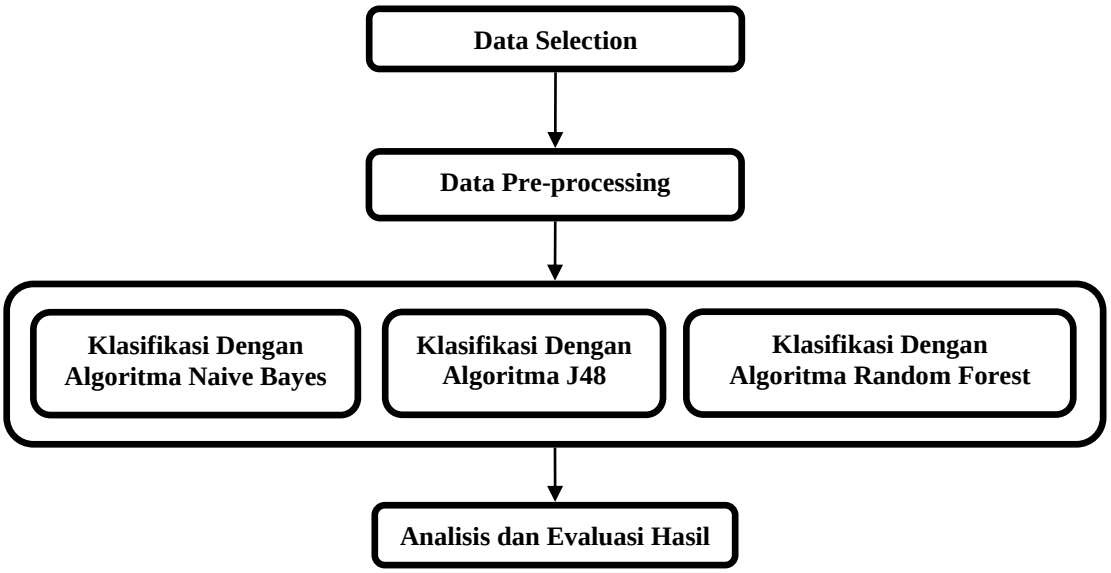
Prediksi lama studi mahasiswa dikembangkan agar dapat mendukung perguruan tinggi dalam membuat kebijakan

agar meningkatkan persentase mahasiswa lulus tepat waktu. Dengan mengetahui hasil prediksi, prodi dibantu oleh dosen pembimbing akademik dapat memberikan perhatian yang lebih terhadap mahasiswa-mahasiswa yang diprediksi tidak lulus tepat waktu, sehingga mereka dapat memperbaiki indeks prestasinya tiap semester (3).

Data mining adalah proses yang menggunakan matematika, teknik statistik, machine learning, dan kecerdasan buatan untuk mengekstraksi dan identifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan di dalam suatu database besar (4). Salah satu teknik yang terdapat dalam data mining adalah teknik klasifikasi. Klasifikasi merupakan suatu proses untuk menemukan model atau fungsi yang menggambarkan class dari sekumpulan data, mendeskripsikan data penting, serta dapat juga memprediksi kecenderungan data di masa depan. Teknik klasifikasi pada penelitian ini menggunakan metode Naive Bayes Classification, J48, dan Random Forest. Ketiga metode tersebut akan dikomparasi pada penelitian untuk mengetahui seberapa tepat akurasi yang diperoleh dari masing-masing metode jika diterapkan pada prediksi lama studi mahasiswa STIKOM Bali

METODOLOGI PENELITIAN

Penelitian ini didasarkan pada penerapan beberapa algoritma klasifikasi yang digunakan untuk memprediksi lama studi mahasiswa di STIKOM Bali. Algoritma klasifikasi yang digunakan adalah Algoritma Naïve Bayes, J48, dan Random Forest yang akan diujikan pada data sampel. Bagan alur proses klasifikasi untuk memprediksi lama studi mahasiswa di STIKOM Bali ditunjukan pada Gambar 1.



Gambar 1. Alur proses klasifikasi untuk prediksi lama studi mahasiswa

Tahap pertama adalah data selection yaitu pemilihan atau seleksi data yang berkaitan dengan objek penelitian. Data yang digunakan penelitian ini berupa data nilai dan data lulusan mahasiswa STIKOM Bali. Data sampel ini didapatkan dari Sistem Informasi Akademik (SINAK) STIKOM Bali. *Data selection* bertujuan untuk membatasi data yang akan digunakan pada penelitian agar tidak terlalu luas dan relevan dengan objek yang diteliti. Hasil dari *data selection* ini selanjutnya digunakan sebagai data uji.

Setelah data sampel terkumpul, langkah berikutnya yaiut menerapkan teknik *pre-processing* data. Data *pre-processing* ini berfungsi untuk membersihkan data, integrasi data, dan transformasi data agar proses klasifikasi lebih optimal. Tahap *pre-processing* ini dapat dilakukan dengan beberapa langkah, seperti menangani data yang tidak lengkap dan tidak konsisten, serta mengurangi data *noise* yang berisi nilai-nilai yang salah dan anomali.

Tahap berikutnya adalah merancang model algoritma dan menyesuaikannya dengan data uji. Algoritma klasifikasi yang digunakan adalah Naive Baye, J48, dan Random Forest. Proses uji coba menggunakan aplikasi WEKA dengan memanfaatkan fitur "Clasify". Hasil uji coba dari masing-masing algoritma kemudian dikomparasi untuk mengetahui algoritma mana yang paling optimal dalam melakukan klasifikasi. Pengukuran akurasi dilakukan dengan menggunakan pendekatan *Correctly Classified Instances*, sedangkan pengukuran error menggunakan *Mean Absolute Error* (MAE).

HASIL DAN PEMBAHASAN

Data Selection dan *Preprocessing*

Data sampel diperoleh langsung dari sistem SINAK pada Bagian Akademik STIKOM Bali, berupa data riwayat nilai mahasiswa lulusan program studi Sistem Komputer dari semester I sampai dengan semester VII. Pemetaan data sampel dilakukan dengan cara menjadikan list matakuliah sebagai atribut data. Dari hasil data selection dan preprocessing diperoleh 41 buah atribut (41 matakuliah), serta record sebanyak 1580 data. Data sampel yang digunakan adalah lulusan mahasiswa reguler (non-transfer) 5 tahun terakhir. Struktur dataset nilai lulusan ditunjukkan seperti pada Tabel 1. Klasifikasi lama studi mahasiswa dibagi menjadi tiga kelas ditunjukkan pada Tabel 2.

Tabel 1. Struktur dataset nilai lulusan

ID Mahasiswa	Matakuliah #1	Matakuliah #2	...	Matakuliah #N	Lama Studi

Tabel 2. Klasifikasi lama studi mahasiswa

CLASS	KETERANGAN
"LEBIH CEPAT"	Masa studi < 4 tahun
"TEPAT WAKTU"	Masa studi = 4 tahun
"LEBIH LAMA"	Masa studi > 4 tahun

Proses berikutnya adalah mengkonversi data sampel yang diperoleh menjadi format ARFF agar dapat diproses lebih lanjut pada aplikasi WEKA. proses konversi ARFF dapat dilakukan dengan menggunakan fitur ARFF Viewer seperti ditunjukan pada Gambar 2.

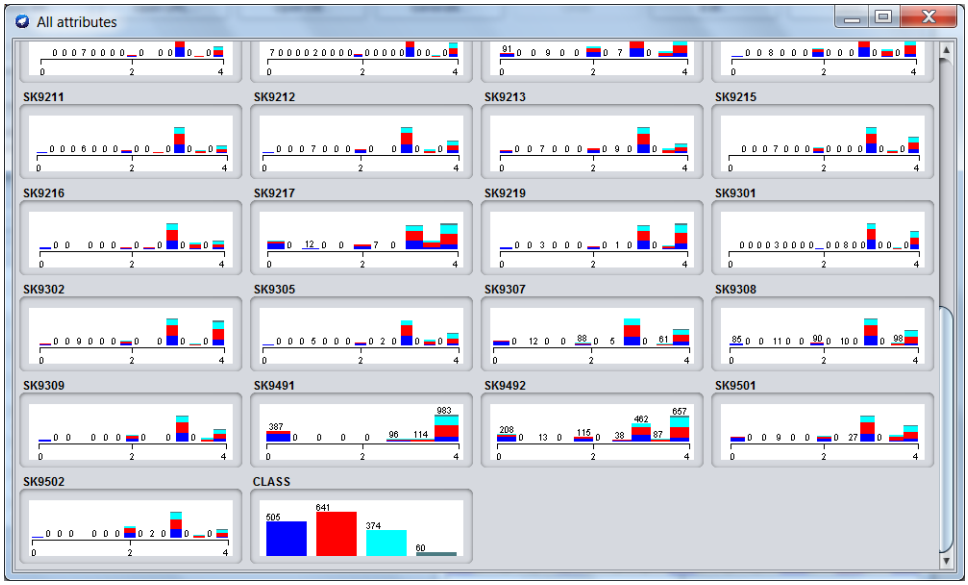
File Edit View

DATA SK.arff

Relation: DATA SK-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,...

No.	1: SK0101 Numeric	2: SK0102 Numeric	3: SK0103 Numeric	4: SK0104 Numeric	5: SK0201 Numeric	6: SK0202 Numeric	7: SK0205 Numeric	8: SK0206 Numeric	9: SK0207 Numeric	10: SK0208 Numeric	11: SK0210 Numeric	12: SK0214 Numeric	13: N
1	4.0	3.0	3.0	3.0	2.0	2.0	3.0	3.0	3.0	4.0	4.0	4.0	
2	4.0	3.0	3.0	2.0	2.0	3.0	1.0	3.0	3.0	2.0	1.0	3.0	
3	4.0	0.0	0.0	0.0	3.0	3.0	3.0	3.0	3.0	3.0	0.0	3.0	
4	3.0	0.0	0.0	0.0	2.0	3.0	2.0	0.0	2.0	0.0	0.0	3.0	
5	3.0	2.0	2.0	3.0	3.0	3.0	3.0	2.0	2.0	2.0	2.0	0.0	
6	3.0	0.0	0.0	0.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	
7	3.0	0.0	0.0	0.0	2.0	2.0	3.0	2.0	3.0	0.0	0.0	3.0	
8	3.0	2.0	2.0	3.0	2.0	3.0	3.0	2.0	3.0	3.0	2.0	3.0	
9	3.0	3.0	3.0	0.0	2.0	3.0	3.0	0.0	2.0	2.0	0.0	3.0	
10	3.0	3.0	3.0	0.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	
11	0.0	2.0	2.0	0.0	2.0	0.0	3.0	0.0	3.0	3.0	3.0	3.0	
12	3.0	3.0	3.0	2.0	2.0	3.0	3.0	0.0	3.0	4.0	3.0	3.0	
13	3.0	3.0	3.0	0.0	3.0	3.0	3.0	2.0	2.0	3.0	0.0	2.0	
14	3.0	3.0	3.0	0.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0	
15	4.0	3.0	3.0	4.0	3.0	4.0	3.0	2.0	3.0	2.0	3.0	3.0	
16	4.0	0.0	0.0	0.0	3.0	3.0	0.0	3.0	3.0	3.0	0.0	3.0	
17	3.0	3.0	3.0	3.0	3.0	3.0	2.0	1.0	0.0	4.0	0.0	1.0	
18	0.0	0.0	0.0	3.0	0.0	0.0	0.0	2.0	0.0	0.0	3.0	3.0	
19	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	
20	3.0	4.0	4.0	0.0	2.0	3.0	0.0	0.0	0.0	3.0	3.0	0.0	
21	0.0	4.0	4.0	3.0	3.0	3.0	0.0	0.0	3.0	3.0	2.0	0.0	

Gambar 2 Data sampel format ARFF



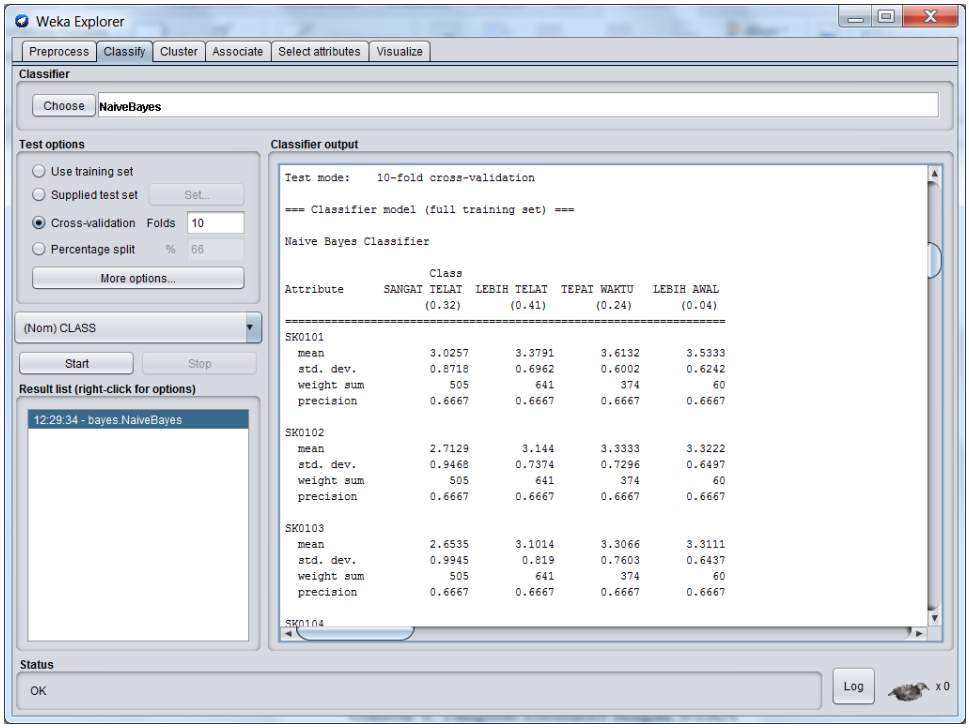
Gambar 3. Visualisasi semua atribut data

PROSES KLASIFIKASI DAN ANALISA HASIL

Proses klasifikasi digunakan untuk memprediksi lama studi mahasiswa pada penelitian ini. Digunakan tiga buah model algoritma klasifikasi, yaitu Naive Bayes, Random Forest, dan J48. Model Naïve Bayes adalah model probabilitas Bayesian yang disederhanakan. Model ini menghitung probabilitas hasil akhir sementara beberapa variabel bukti terkait diberikan. Probabilitas variabel bukti diasumsikan independen terhadap probabilitas variabel bukti lainnya, karena hasil akhir yang sama terjadi. Pada tahap pelatihan, algoritme Naïve Bayes menghitung probabilitas dari hasil yang diberikan untuk atribut tertentu dan kemudian menyimpan probabilitas ini. Proses ini dilanjutkan untuk setiap atribut. Pada tahap pengujian, jumlah waktu yang dibutuhkan untuk menghitung probabilitas kelas yang diberikan untuk setiap contoh dalam kasus terburuk sebanding dengan N, jumlah atribut (5).

Metode Random Forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (*bagging*) dan random feature selection. Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas N amatan dan P peubah penjelas.

J48 (C4.5) merupakan salah satu algoritma pohon keputusan pengembangan dari ID3. Pohon dibentuk dengan membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama. Split yang digunakan untuk membagi data tergantung dari jenis atribut yang digunakan. Atribut *numeric* dipecah dengan mengurutkan sampel berdasarkan atribut kontinyu A, kemudian membentuk *minimum threshold* M. Atribut diskret memputnyai bentuk pemecahan value(A) ε X, dimana X ⊂ domain(A) (6).



Gambar 4. Tampilan klasifikasi dengan WEKA

Proses klasifikasi menggunakan fitur "Clasify" pada aplikasi WEKA. evaluasi kinerja model menggunakan *10 folds cross-valdiation*, dimana data sampel dipisahkan menjadi dua subset dengan komposisi 9/10 sebagai *data training* dan 1/10 sebagai *data testing*. Tampilan proses klasifikasi dengan aplikasi WEKA ditunjukkan pada Gambar 4. Rangkuman hasil uji coba klasifikasi dengan algoritma Naïve Bayes (NB), Random Forest (RF), dan J48 ditunjukan pada Tabel 3.

Tabel 3. Hasil uji coba proses klasifikasi NB, RF, dan J48

	NB	**RF**	**J48**
Correctly Classified Instances	69.96%	77.99%	74.86%
Incorrectly Classified Instances	30.04%	22.01%	25.14%
Kappa statistic	0.3785	0.3441	0.2885
Mean absolute error	0.2066	0.223	0.2276
Root mean squared error	0.4168	0.3216	0.3693
Relative absolute error	74.09%	79.95%	81.61%
Root relative squared error	111.71%	86.20%	98.98%
Time to build model	0.01 sec	0.31 sec	0.03 sec

Berdasarkan hasil uji coba diatas, tingkat akurasi dari masing-masing algoritma ditunjukan pada bagian "Correctly Classified Instances". Akurasi tertinggi diperoleh dengan menggunakan algoritma Random Forest yaitu 77.99% dan paling rendah diperoleh dengan metode Naive Bayes yaitu 69.96%. untuk pengukuran tingkat kesalahan/error menggunakan pendekatan "Mean absolute error" atau MAE. Tingkat kesalahan paling kecil dihasilkan oleh algoritma Naïve Bayes dengan nilai MAE = 0.2066, sedangkan tingkat error terbesar dihasilkan oleh J48

dengan nilai MAE = 0.2276. dari hasil uji coba diatas menunjukan ketiga algoritma yang diujikan telah mampu melakukan klasifikasi data lulusan untuk prediksi lama studi mahasiswa dengan rata-rata akurasi diatas 60%..

KESIMPULAN

Algoritma klasifikasi yang diuji coba telah mampu melakukan prediksi lama studi mahasiswa. Data sampel diperoleh langsung dari sistem SINAK pada bagian Akademik STIKOM Bali. Diperoleh 1580 data dan 41 matakuliah (atribut data) pada tahap *data selection* dan *preprocessing*. Model klasifikasi menggunakan algoritma Naïve Bayes, Random Forest dan J48. Uji coba menggunakan aplikasi WEKA, dengan evaluasi kinerja model menggunakan 10 *folds cross-validation*. Berdasarkan hasil uji coba menunjukkan algoritma Random Forest menghasilkan akurasi terbesar yaitu 77.99% dan paling rendah diperoleh dengan metode Naive Bayes yaitu 69.96%. Sedangkan tingkat kesalahan paling kecil dihasilkan oleh algoritma Naïve Bayes dengan nilai MAE = 0.2066 dan terbesar dihasilkan oleh J48 dengan nilai MAE = 0.2276

DAFTAR PUSTAKA

1. Marquez-Vera C, Romero C, Ventura S. Predicting School Failure Using Data Mining. Proc 4th Int Conf Educ Data Min [Internet]. 2011;(December):271–6. Available from: http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper11_short_Marquez-Vera.pdf
2. Baker RSJD, Yacef K. The State of Educational Data Mining in 2009 : A Review and Future Visions. J Educ Data Min. 2009;1(1):3–16.
3. Saputra AY, Primadasa Y. Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. TechnoCom. 2018;17(4):395–403.
4. Turban E, Aronson JE, Liang T-P. Decision Support Systems and Intelligent Systems. Yogyakarta: Andi Offset; 2005.
5. Mongkareng D, Setiawan NA, Permanasari AE. Implementasi Data Mining dengan Seleksi Fitur untuk Klasifikasi Serangan pada Intrusion Detection System (IDS). Citee. 2017;(gambar 2):314–21.
6. Astuti T, Mujiati I, Ayu D, Ristianah V, Lestari WA. Penerapan Algoritme J48 Untuk Prediksi. J Telemat. 2016;9(2):1–10.