PREDIKSI LAMA STUDI MAHASISWA DENGAN METODE RANDOM FOREST (STUDI KASUS : STIKOM BALI)

Prediction Of Students' Learning Study Periode By Using Random Forest Method (Case Study: Stikom Bali)

I Made Budi Adnyana
STMIK STIKOM Bali
Jln. Raya Puputan No.86, Renon, Denpasar
e-mail: budi.adnyana@stikom-bali.ac.id

Abstrak

Kelulusan tepat waktu merupakan salah satu elemen penilaian akreditasi dari perguruan tinggi. Selain itu wisuda tepat waktu merupakan isu yang penting karena tingkat kelulusan sebagai dasar efektifnya suatu perguruan tinggi. Kurangnya informasi dan analisa yang diperoleh Bidang Akademik STIKOM Bali mengakibatkan sulitnya melakukan prediksi lama studi mahasiswa. Prediksi lama studi mahasiswa dapat membantu Bidang Akademik dalam menyusun strategi yang tepat untuk menekan atau memperpendek lama studi mahasiswa. Pada permasalahan ini dapat diterapkan teknik data mining untuk melakukan prediksi yaitu dengan menggunakan metode klasifikasi Random Forest. Random Forest merupakan suatu kumpulan dari beberapa tree, dimana masing-masing tree bergantung pada nilai piksel pada tiap vektor yang daimbil secara acak dan independen. Data sampel diperoleh langsung dari bagian Akademik STIKOM Bali. Data yang digunakan adalah data lulusan 2 tahun terakhir, meliputi IPK, SKS, jumlah cuti dan non-aktif, nilai mahasiswa, dan lama studi mahasiswa. Output dari sistem ini berupa klasifikasi yang terdiri dari 2 kelas, yaitu lulus tepat waktu dan lulus lewat batas waktu. Dari hasil eksperimen diperoleh nilai akurasi adalah 83.54%.

Kata kunci: *prediksi, lama studi, random forest*

Abstract

Graduation on time is one of the assessment elements of the college accreditation. Furthermore, graduation on time is an important issue because it indicates an effectiveness of college. Academic division of STIKOM Bali face many difficulties on predicting student graduation time because lack of information and analysis. Predictions of graduation time can help academic division in making appropriate strategy to shorten the study time. Data mining can be applied on this prediction problems using random forest classification methods. Random forest is a collection of of several tree, where each tree dependent on the pixels on each vector that selected randomly and independent .Sample data obtained from academic division of STIKOM Bali. This research use sample data of last 2 years graduated students, such as IPK, SKS, the number of inactive, and study time. The classification output consists of 2 class, "graduate on time" and "graduate over the time". From the experimental result, 83.54 % accuracy value obtained.

Keywords*: prediction, study time, random forest*

1. PENDAHULUAN

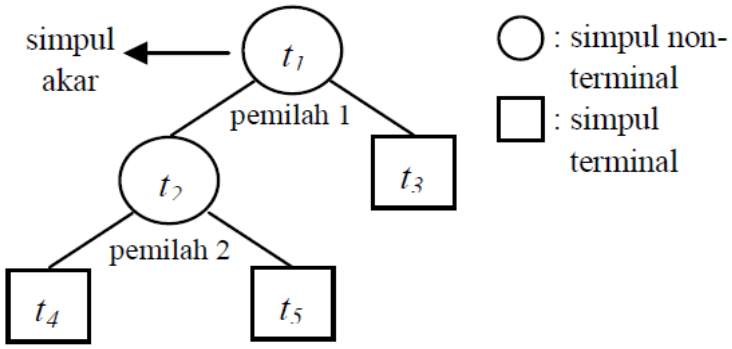
STIKOM Bali merupakan salah satu perguruan tinggi TIK yang terdapat di Denpasar yang berdiri sejak tahun 2002. STIKOM Bali memiliki tiga buah Program Studi yaitu Sistem Komputer (S1), Sistem Informasi (S1), dan Manajemen Informatika (D3). Jumlah mahasiswa dan dosen yang dimiliki berdasarkan data dari PDDIKTI adalah 4.203 orang mahasiswa dan 151 orang dosen pada tahun awal 2016, dengan rasio dosen 1 : 27.8.

Saat ini intitusi perguruan tinggi berada dalam lingkungan yang sangat kompetitif. Setiap perguruan tinggi berusaha untuk terus memperbaiki menejemennya untuk meningkatkan mutu pendidikan dan meningkatkan akreditasi. Salah satu elemen penilaian akreditasi perguruan tinggi adalah lulus tepat waktu. Selain itu wisuda tepat waktu merupakan isu yang penting karena tingkat kelulusan sebagai dasar efektifnya suatu perguruan tinggi [1].

Seiring dengan perkembangan jumlah mahasiswa di STIKOM Bali, hal penting yang harus diperhatikan juga adalah kelulusan mahasiswa. Jumlah total lulusan di STIKOM Bali telah mencapai 3.601 mahasiswa. Rata-rata lama studi mahasiswa yang lulus pada wisuda XVI adalah 4.8 tahun untuk mahasiswa S1 dan 3.9 tahun untuk mahasiswa D3. Hal ini menunjukkan rata-rata lama studi mahasiswa di STIKOM Bali lebih lama atau melampaui masa studi standar. Hal ini dapat disebabkan oleh banyak faktor, seperti keaktifan mahasiswa, nilai studi mahasiswa, faktor biaya, dan sebagainya. Kurangnya informasi dan analisa yang diperoleh Bidang Akademik mengakibatkan sulitnya melakukan prediksi lama studi mahasiswa. Prediksi lama studi mahasiswa dapat membantu Bidang Akademik dalam menyusun strategi yang tepat untuk menekan atau memperpendek lama studi mahasiswa.

Diperlukan suatu solusi untuk dapat mengatasi permasalahan diatas, yaitu sebuah model yang dapat membantu Bidang Akademik STIKOM Bali untuk melakukan prediksi lama studi mahasiswanya. Pada permasalahan ini dapat diterapkan teknik data mining untuk melakukan prediksi yaitu dengan menggunakan metode klasifikasi Random Forest. Random forest merupakan pengembangan dari Decision Tree dengan menggunakan beberapa Decision Tree, dimana setiap Decision Tree telah dilakukan training menggunakan sampel individu dan setiap atribut dipecah pada tree yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada vote dari suara terbanyak pada kumpulan populasi tree. Metode klasifikasi Random Forest ini telah diaplikasikan pada berbagai permasalahan seperti prediksi mahasiswa berpotensi non-aktif [1], permasalahan *driver analysis* [2], dan permasalahan klasifikasi hutan-non hutan [3].

CART (Classification and Regression Tree) merupakan metode eksplorasi data yang didasarkan pada teknik pohon keputusan. Pohon klasifikasi dihasilkan saat peubah respons berupa data kategorik, sedangkan pohon regresi dihasilkan saat peubah respons berupa data numerik. Pohon terbentuk dari proses pemilahan rekursif biner pada suatu gugus data sehingga nilai peubah respons pada setiap gugus data hasil pemilahan akan lebih.



Gambar 1 Struktur Pohon pada Metode CART.

Pohon diilustrasikan dalam Gambar 1. Pohon disusun oleh simpul t1, t2, …, t5 (Gambar 1). Setiap pemilah (*split*) memilah simpul non-terminal menjadi dua simpul yang saling lepas. Hasil prediksi respons suatu amatan terdapat pada simpul terminal. Pembangunan pohon klasifikasi CART meliputi tiga hal, yaitu:

1. Pemilihan pemilah (*split*)
2. Penentuan simpul terminal
3. Penandaan label kelas

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, random forest dilakukan dengan cara [4]:

a) Lakukan penarikan contoh acak berukuran *n* dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
b) Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari *m* peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
c) Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

Respons suatu amatan diprediksi dengan menggabungkan (*aggregating*) hasil prediksi k pohon. Pada masalah klasifikasi dilakukan berdasarkan majority vote (suara terbanyak). Error klasifikasi random forest diduga melalui error OOB yang diperoleh dengan cara [5]:

a) Lakukan prediksi terhadap setiap data OOB pada pohon yang bersesuaian. Data OOB (out of bag) adalah data yang tidak termuat dalam contoh bootstrap.
b) Secara rata-rata, setiap amatan gugus data asli akan menjadi data OOB sebanyak sekitar 36% dari banyak pohon. Oleh karena itu, pada langkah 1, masing-masing amatan gugus data asli mengalami prediksi sebanyak sekitar sepertiga kali dari banyaknya pohon. Jika a adalah sebuah amatan dari gugus data asli, maka hasil prediksi random forest terhadap a adalah gabungan dari hasil prediksi setiap kali a menjadi data OOB.
c) Error OOB dihitung dari proporsi misklasifikasi hasil prediksi random forest dari seluruh amatan gugus data asli.

Breiman dan Cutler (2005) menyarankan untuk mengamati error OOB saat *k* kecil, lalu memilih m yang menghasilkan error OOB terkecil. Jika random forest dilakukan dengan menghasilkan variable importance, disarankan untuk menggunakan banyak pohon, misalnya 1000 pohon atau lebih. Jika peubah penjelas yang dianalisis sangat banyak, nilai tersebut dapat lebih besar agar *variable importance* yang dihasilkan semakin stabil. [6].

Dengan diterapkannya metode Random Forest ini pada permasalahan data mining untuk prediksi lama studi mahasiswa, diharapkan dapat membantu Bidang Akademik dalam menyusun strategi yang tepat untuk menekan dan mengurangi rata-rata lama studi mahasiswa di STIKOM Bali untuk periode wisuda berikutnya.

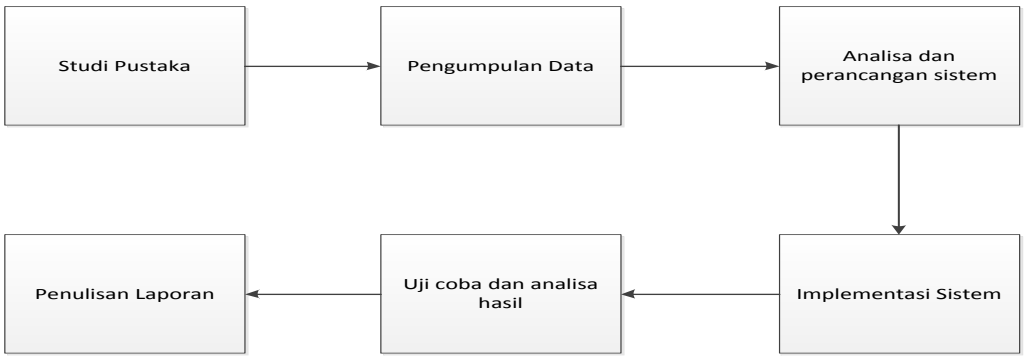
2. METODE PENELITIAN

Bab ini berisi tentang metode penelitian dan cara pendekatan yang digunakan pada penelitian, serta sistematika penelitian yang secara umum dijelaskan sebagai berikut.

2.1. Tempat dan Waktu Penelitian

Penelitian yang berjudul "Data Mining Untuk Memprediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : STMIK STIKOM Bali)" ini dilakukan di STIKOM Bali. Penelitian ini dilaksanakan selama 5 bulan.

2.2. Sistematika Penelitian



Gambar 2 Bagan sistematika penelitian

2.3. Pengumpulan Data

Berdasarkan sumbernya, data penelitian dapat dikelompokkan menjadi dua jenis yaitu data primer dan data sekunder. Pemahaman terhadap kedua jenis data tersebut diperlukan sebagai landasan dalam menentukan teknik serta langkahlangkah pengumpulan data penelitian.Data Primer adalah data yang diperoleh atau dikumpulkan oleh peneliti secara langsung dari sumber datanya. Data primer disebut juga sebagai data asli atau data baru yang memiliki sifat up to date. Untuk mendapatkan data primer, peneliti harus mengumpulkannya secara langsung. Teknik yang digunakan peneliti untuk mengumpulkan data primer antara lain observasi, wawancara, dan diskusi terfokus. Data Sekunder adalah data yang diperoleh atau dikumpulkan peneliti dari berbagai sumber yang telah ada (peneliti sebagai tangan kedua). Data sekunder dapat diperoleh dari berbagai sumber seperti buku, laporan, jurnal, dan lain-lain. Data primer yang digunakan dalam penelitian ini diperoleh langsung dari sumber datanya yaitu pada Bidang Akademik STIKOM Bali dengan beberapa macam teknik, seperti observasi, wawancara, serta diskusi terfokus. Data yang diperlukan dalam penelitian yang diusulkan ini adalah berupa data riwayat kelulusan mahasiswa, riwayat keaktifan mahasiswa, dan sebagainya yang berpengaruh terhadap lama studi mahasiswa.

2.4. Analisa dan Perancangan Sistem

Pada tahap ini dilakukan analisa terhadap data yang sudah terkumpul dan analisa terhadap alur proses sistem prediksi menggunakan metode random forest. Perancangan serta implementasi metode random forest pada sistem prediksi lama studi ini akan digambarkan dalam bentuk bagan flowchart.

2.5. Implementasi dan Uji Coba

Pada tahap ini sistem yang sudah dikembangkan diimplementasikan atau diuji coba pada permasalahan yang dihadapi, yaitu permasalahan prediksi lama studi mahasiswa dengan menggunakan data yang diperoleh langsung di STIKOM Bali. Hasil uji coba akan dianalisa lebih lanjut sehingga dapat ditarik kesimpulan pada penelitian ini.

3. HASIL DAN PEMBAHASAN

Hasil penelitian yang akan dibahas pada bab ini secara garis besar meliputi pengolahan awal data kelulusan mahasiswa dari STIKOM Bali, pemrosesan data mining menggunakan metode Random Forest, dan pembahasan hasil uji coba.

3.1. Data Penelitian

Data yang digunakan pada penelitian ini diperoleh langsung dari bagian Akademik STIKOM Bali. Untuk melakukan prediksi lama studi mahasiswa, pada penelitian ini menggunakan history data mahasiswa yang sudah lulus 2 tahun terakhir, yaitu kelulusan periode tahun 2014/2015 dan 2015/2016. Adapun beberapa fitur data mahasiswa yang digunakan sebagai input pada sistem adalah:

a) IPK (pada semester VI)
b) Jumlah total SKS (pada semester VI)
c) Jumlah tidak aktif dan cuti (sampai semester VI)
d) Jumlah matakuliah dengan nilai buruk (nilai D dan E)
e) Jumlah matakuliah dengan nilai bagus (nilai A dan B)

Data fitur yang digunakan tersebut diambil pada atau sampai dengan semester VI dari mahasiswa bersangkutan. Hal ini bertujuan untuk mengetahui pengaruh ipk, sks, keaktifan serta nilai mahasiswa yang sudah ditempuh sampai semester VI terhadap waktu kelulusan mahasiswa tersebut. Adapun fitur data yang akan dijadikan kelas atau output dari sistem adalah:

a) Lama studi mahasiswa bersangkutan

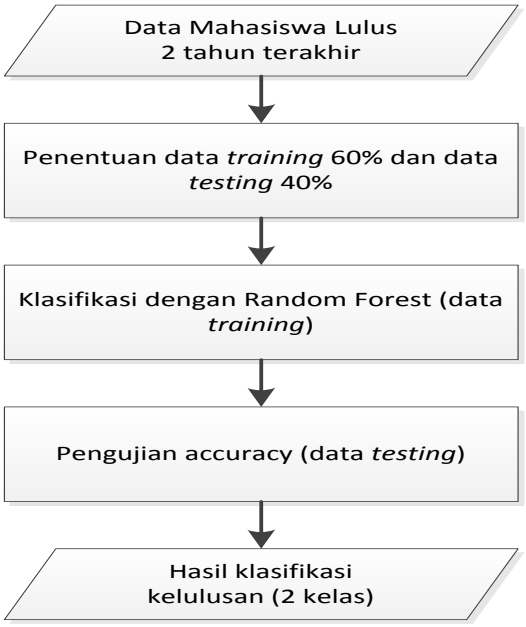
b) Kelas yang dibagi menjadi 2 yaitu: "lulus tepat waktu" dan "lulus lewat batas waktu". Dikategorikan lulus tepat waktu jika lama studi lebih kecil atau sama dengan 4 tahun bagi mahasiswa jenjang S1 dan "lulus lewat batas waktu" jika lama studi lebih dari 4 tahun.

Data uji yang digunakan pada penelitian ini adalah data mahasiswa dari jenjang S1 dan mahasiswa non-transfer saja, dimana masa studi standarnya adalah 4 tahun atau 8 semester. Pemilihan kriteria data uji ini bertujuan untuk membatasi lingkup data agar lebih mudah untuk dilakukan proses klasifikasi. Contoh data yang diperoleh di STIKOM Bali ditunjukkan pada Tabel 1.

Tabel 1 Contoh data sampel mahasiswa lulusan STIKOM Bali (2 tahun terakhir)

IPK	SKS	Jumlah Cuti/Non-Aktif	Jumlah Mk Nilai Buruk	Jumlah Mk Nilai Bagus	Lama Studi	Keterangan
3.62	144	0	0	84	4	Tepat waktu
3.67	146	1	2	94	4	Tepat waktu
3.13	140	1	11	80	4.5	Lewat batas waktu
3.57	146	1	0	94	4	Tepat waktu
3.57	130	1	6	85	4.5	Lewat batas waktu
3.14	128	1	8	83	4.5	Lewat batas waktu
3.27	142	0	1	76	4	Tepat waktu
2.71	126	0	23	68	5	Lewat batas waktu
2.98	118	1	11	66	4.5	Lewat batas waktu
3.50	142	0	2	77	4	Tepat waktu
3.41	138	1	9	84	4	Tepat waktu
3.57	136	0	1	75	4	Tepat waktu

Bagan umum dari alur kerja metode Random Forest dalam mengolah data sehingga dihasilkannya model prediksi lama studi mahasiswa di STIKOM Bali ditunjukkan pada Gambar 3 berikut ini.

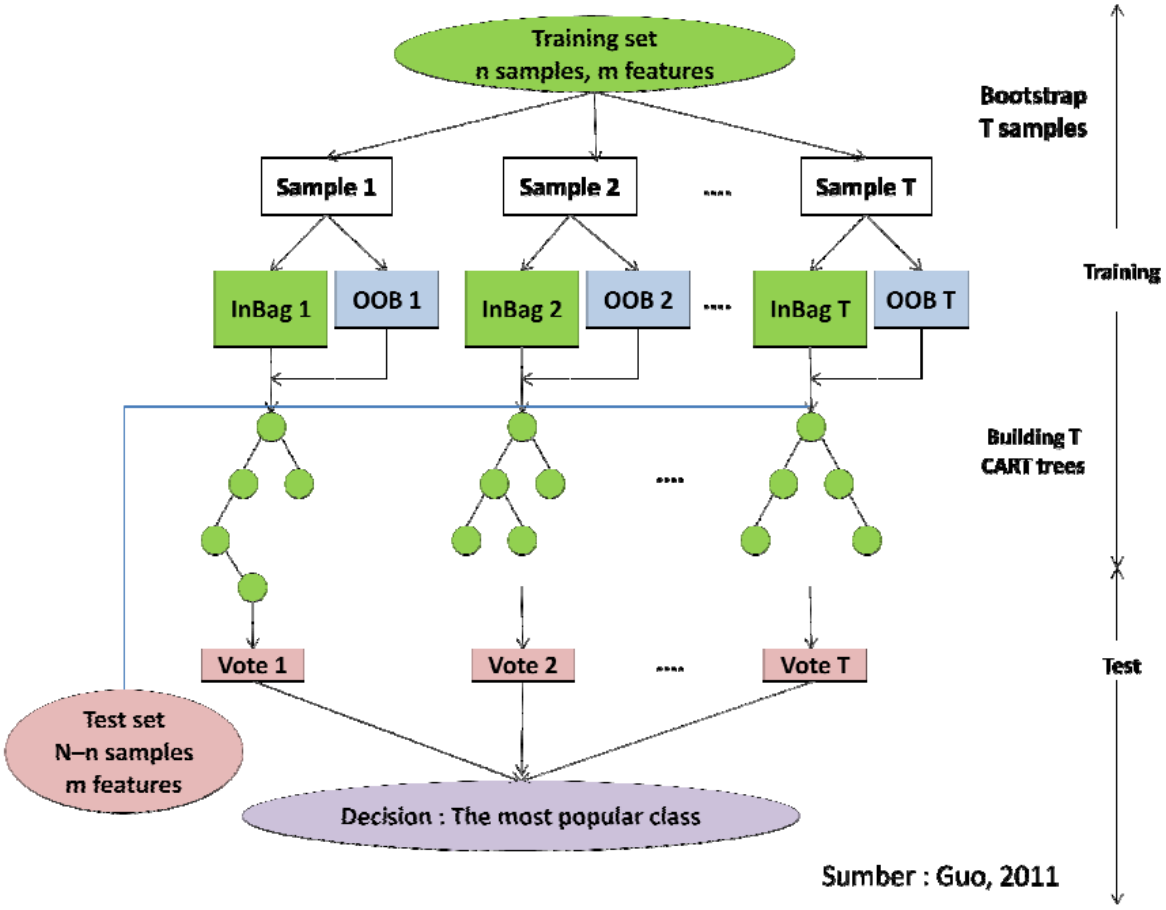


Gambar 3 Bagan alur prediksi lama studi mahasiswa dengan Random Forest

3.2. Proses klasifikasi dengan metode Random Forest

Proses prediksi kelulusan dilakukan dengan metode Random Forest non-parametrik. Random Forest merupakan suatu kumpulan dari beberapa tree, dimana masing-masing tree bergantung pada nilai piksel pada tiap vektor yang daimbil secara acak dan independen. Random Forest tidak berkecenderungan untuk overfit dan dapat melakukan proses dengan cepat, sehingga memungkinkan untuk memproses tree sebanyak yang diinginkan oleh pengguna[3].

Gambar 4 menampilkan alur prediksi menggunakan metode Random Forest. Dalam pembentukan tree, algoritma Random Forest akan melakukan training terhadap sampel data. Data training yang digunakan adalah 60% dari sampel data, sisanya 40% digunakan sebagai data uji. Sebanyak sepertiga sampel data dari training set digunakan sebagai data out of bag (OOB). Data OOB digunakan untuk menghitung error dan menentukan variable importance. Variabel yang digunakan untuk pemisahan (split) terbaik ditentukan secara acak. Setelah seluruh tree terbentuk, maka proses klasifikasi akan berjalan. Penentuan kelas dilakukan dengan cara voting dari masing-masing tree, kelas dengan kelas terbanyak akan menjadi pemenangnya [5]



Gambar 4 Proses klasifikasi Random Forest

3.3. Hasil Pengujian

Tools yang digunakan untuk melakukan proses Random Forest adalah aplikasi WEKA. Hasil pengujian menunjukkan metode Random Forest selesai dengan menggunakan 10 buah *tree*, masing-masing dibangun dengan mempertimbangkan 3 buah fitur secara acak. Hasil estimasi dari out of bag error = 0.1677. Hasil evaluasi dari *test split* ditunjukkan pada Tabel 2.

Tabel 2 Hasil evaluasi *test split*

Correctly Classified Instances	269 (83.54%)
Incorrectly Classified Instances	53 (16.45%)
Kappa statistic	0.564
Mean absolute error	0.1848
Root mean squared error	0.3116
Relative absolute error	50.45%
Root relative squared error	75.35%
Total Number of Instances	322

Hasil pengukuran akurasi dari model yang dikembangkan ditampilkan pada Tabel 3.

Tabel 3 Hasil pengukuran akurasi Random Forest

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
TepatWaktu	0.783	0.15	0.587	0.783	0.671	0.913
LewatBatasWaktu	0.85	0.217	0.935	0.85	0.89	0.913
Rata-rata	0.835	0.203	0.86	0.835	0.843	0.913

Confussion matrix ditampilkan dalam Tabel 4. Untuk kelas "Lulus Tepat Waktu" diperoleh nilai *producer accuracy* sebesar 81,81 % dan kelas "Lulus Lewat Waktu Studi" diperoleh nilai *producer accuracy* sebesar 84,98%. Sedangkan *overall accuracy* yang didapatkan adalah 83.54% (Kappa Coefficient 0.564).

Tabel 4 *Confussion Matrix* hasil klasifikasi prediksi lama studi

Data Referensi / **Hasil Klasifikasi**	**Tepat Waktu**	**Lewat Batas Waktu**	***User's Accuracy***
Tepat Waktu	54	38	58.69%
Lewat Batas Waktu	15	215	93.48%
Producers's Accuracy	81.81%	84.98%	

4. KESIMPULAN

Berdasarkan hasil dan pembahasan pada penelitian ini, maka dapat disimpulkan beberapa hal sebagai berikut:

a. Terdapat 5 buah fitur data yang digunakan sebagai input dari sistem dan output berupa prediksi lama studi atau klasifikasi yang terdiri dari 2 kelas, yaitu "lulus tepat waktu" dan "lulus lewat batas waktu studi".
b. Hasil pengujian menunjukkan nilai *producer accuracy* untuk kelas "Lulus Tepat Waktu" diperoleh nilai sebesar 81,81 % dan kelas "Lulus Lewat Waktu Studi" diperoleh nilai *producer accuracy* sebesar 84,98%. Sedangkan *overall accuracy* yang didapatkan adalah 83.54% (Kappa Coefficient 0.564).

5. SARAN

Untuk lebih meningkatkan kualitas hasil klasifikasi, maka dalam penelitian selanjutnya dapat diterapkan algoritma *pre-processing* dan *feature selection* terhadap data sampel.

DAFTAR PUSTAKA

[1] Dwi Untari, 2014. *Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5*. Program Studi Teknik Informatika - Fakultas Ilmu Komputer - Universitas Dian Nuswantoro

[2] Dewi, N.K., Syafitri, U.D., Mulyadi, S.Y. 2011. *Penerapan Metode Random Forest Dalam Driver Analysis (The Application of Random Forest in Driver Analysis)*, Forum Statistika dan Komputasi. 16, 35 – 43

[3] Sambodo, K.A., Rahayu, M.I., Indriasari, N., Natsir, M., 2014. *Klasifikasi Hutan-Non Hutan Data Alos Palsar Menggunakan Metode Random Forest*, Seminar Nasional Penginderaan Jauh, 120 – 127

[4] Breiman L. 1996. *Bagging Predictors*. Machine Learning 24, 123-140

[5] Breiman L. 2001. *Random Forests*. Machine Learning 45, 5-32

[6] Breiman, L., and A. Cutler. 2005. Random Forests http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (diakses tgl 27 Juli 2016).