Prediksi Ketepatan Masa Studi Mahasiswa dengan Algoritma Pohon Keputusan C45

Studi Kasus: Teknik Informatika UII

Sri Yunianita, Novi Setiani[1], Sri Mulyati
Jurusan Teknik Informatika Fakultas Teknologi Industri
Universitas Islam Indonesia
Yogyakarta
[1]novi.setiani@uii.ac.id

***Abstrak*—Setiap program studi harus melakukan evaluasi diri sebagai upaya untuk mengetahui keadaan dan peluang-peluang yang dimiliki dan dihadapi secara objektif. Salah satu komponen evaluasi yang perlu diukur adalah kurun waktu penyelesaian studi mahasiswa. Sementara itu ditemukan fakta bahwa kelulusan tepat waktu mahasiswa Program Studi Teknik Informatika UII angkatan 2009 sampai 2011 mengalami penurunan. Oleh karena itu, diperlukan upaya untuk menangani masalah ini salah satunya melalui prediksi terhadap ketepatan masa studi mahasiswa. Hasil prediksi ini dapat membantu pengelola program studi dalam membina mahasiswanya melalui pengelompokkan berdasarkan ketepatan masa studi. Prediksi ini memanfaatkan data alumni untuk digunakan sebagai data latih dalam tahap pemodelan prediksi kelulusan mahasiswa. Seluruh atribut data latih diseleksi dengan algoritma *feature selection* dan diprediksi menggunakan algoritma C4.5. Hasil eksperimen yang memberikan nilai akurasi sebesar 73.99% menunjukkan bahwa model ini dapat digunakan dan masih perlu dikembangkan lagi dengan menggunakan teknik lainnya.**

Kata kunci—prediksi masa studi; c4.5; pohon keputusan; feature selection

I. PENDAHULUAN

Menurut [1] hasil akademik mahasiswa merupakan suatu hal yang krusial bagi institusi pendidikan karena hal tersebut memungkinkan implementasi program yang bertujuan untuk perbaikan dan pemeliharaan hasil akademik selama mahasiswa menuntut ilmu di perguruan tinggi tersebut.

Mahasiswa dan lulusan memiliki peranan penting dalam daur penjaminan mutu program studi. Seperti dalam proses evaluasi diri, salah satu komponennya adalah mahasiswa dan lulusan dengan hasil pembelajaran berupa data tentang kemajuan, keberhasilan, dan kurun waktu penyelesaian studi mahasiswa (termasuk IPK dan yudisium lulusan). Selain itu, mahasiswa dan lulusan menjadi muatan salah satu standar akreditasi yang beberapa deskripsi elemen penilaiannya adalah jumlah mahasiswa dan lulusan 7 (tujuh) tahun terakhir untuk jenjang S-1 disertai rata-rata masa studi dan IPK lulusan 3 (tiga) tahun terakhir.

Oleh karena itu diperlukan pemantauan dan evaluasi terkait tingkat kelulusan mahasiswa secara berkala dengan menemukan pengetahuan atau pola kecenderungan tingkat kelulusan mahasiswa berdasarkan capaian akademik. Pengetahuan ini dapat digunakan sebagai bahan pertimbangan Program Studi untuk menentukan kebijakan atau pembinaan terkait tingkat kelulusan mahasiswanya.

Algoritma yang digunakan pada prediksi masa studi mahasiswa ini adalah algoritma pohon keputusan C4.5. Algoritma C4.5 adalah salah satu algoritma pada metode klasifikasi *data mining* yang memanfaatkan data latih yang bersifat kontinyu maupun diskrit untuk merepresentasikan aturan. Algoritma ini mampu menangani *data training* yang mengandung *missing value* dan memangkas aturan yang tidak diperlukan. Popularitas dari keunggulannya membuat algoritma C4.5 menempati urutan pertama dalam 10 algoritma paling popular menurut *6th International Conference on Data Mining* [2].

Makalah ini ditulis dalam lima bagian, bagian pertama adalah Bab I yang berisi pendahuluan mengenai latar belakang masalah, motivasi dan usulan solusi. Pada Bab II diuraikan mengenai penelitian-penelitian yang terkait dengan pemanfaatan teknik data mining untuk prediksi masa studi mahasiswa. Pada Bab III dijelaskan mengenai tahapan dalam melakukan prediksi dengan menggunakan pendekatan CRISP DM [3] . Pada Bab IV diuraikan hasil pengujian model prediksi, dan pengambilan simpulan dijelaskan pada Bab V.

II. PENELITIAN TERKAIT

Penelitian mengenai pemanfaatan teknik *data mining* untuk memprediksi performansi mahasiswa sudah banyak dilakukan. Hal ini merupakan salah satu implikasi dari meningkatnya volume data akademik yang dikelola secara digital. [4] telah melakukan *review* terhadap makalah yang dipublikasikan mulai tahun 2002 sampai 2014 mengenai prediksi performansi mahasiswa yang menggunakan teknik data mining seperti *k-Nearest neighbor*, *Naïve bayes*, *SVM*, pohon keputusan dan *Neural Network*. Salah satu hasil dari review tersebut adalah algoritma pohon keputusan memberikan hasil yang cukup baik untuk memprediksi performansi peserta didik dengan memanfaatkan atribut penilaian internal seperti nilai kuis, nilai tugas dan nilai ujian.

Dalam konteks penelitian ini, performansi mahasiswa diukur melalui lama masa studi. Beberapa penelitian telah dilakukan untuk memprediksi lama masa studi mahasiswa di beberapa program studi di Indonesia. Pada Tabel I dirangkum beberapa penelitian dalam topik prediksi kelulusan mahasiswa yang dipubikasikan mulai tahun 2015 sampai 2017.

TABEL I. PERBANDINGAN PENELITIAN SEJENIS

No	Atribut prediksi	Seleksi fitur	Algoritma	Penulis
1.	IPK dan Jurusan	Ya	SVM	Damanik, et.al [5]
2.	Jenis kelamin, Program studi, lama skripsi, IPK, nilai ujian masuk	Tidak	Chaid regression tree	Suniantara, dan Rusli[6]
3.	Nilai mata kuliah	Tidak	Naïve bayes	Sari, K [7]
4.	IPK Semester 4 dan status pekerjaan orang tua	Ya	Naïve bayes	Nugroho [8]
5.	Demografi mahasiswa dan IPK	Tidak	k-Nearest neighbor dan Neural network	Prasetyo, T.F. [9]
6.	Nilai mata kuliah	Tidak	Decision tree	Gunawan [10]
7.	Fakultas, jenis kelamin, usia dan IP per semester	Tidak	k-Nearest neighbor	Rohman, A.[11]

III. METODE

Implementasi teknik data mining pada penelitian ini mengikuti tahapan dalam metode CRISP DM [3] yang diuraikan sebagai berikut.

A. Business Understanding Phase

Adapun tujuan penelitian ini adalah untuk menggali pengetahuan (*discovering knowledge*) mengenai pemodelan aturan untuk memprediksi kelulusan mahasiswa apakah tergolong dalam klasifikasi tepat atau tidak tepat berdasarkan data yang dimiliki mahasiswa.

B. Data Understanding Phase

Data latih dan data uji yang digunakan adalah data mahasiswa 2010 sampai 2013. Atribut prediktif yang digunakan adalah nilai mata kuliah wajib yang sudah ditempuh dan data demografi mahasiswa. Atribut kelas yang akan diprediksi adalah apakah mahasiswa tersebut lulus dalam kurun waktu kurang/sama dengan empat tahun (tepat waktu) atau lebih dari empat tahun (tidak tepat waktu).

Data yang digunakan dalam penelitian ini diperoleh dari basisdata Badan Sistem Informasi UII yang terdiri dari dua dokumen excel. Dokumen pertama berisi 82.477 baris data dan 9 kolom yaitu nim, id_matkul, matakuliah, sks, th_akademi, smt, nilai, kd_jurusan, dan kd_kurikulum. Sedangkan dokumen kedua berisi 1.251 baris data dan 13 kolom yaitu nim, ipk, jumlah_sks, email_mhs, kelamin, nip_dpa, dpa, provinsi, jalur, tahun_lulus, smt_lulus, tgl_yudisium, dan tgl_pendadaran.

Tahap eksplorasi data dilakukan dengan terlebih dahulu menyimpan data dari kedua dokumen ke *database* dengan nama *database* spk. Data dari dokumen pertama diunggah ke tabel data_train1, sedangkan data dari dokumen kedua diunggah ke tabel data _train2. Penelusuran terhadap data pada tabel data_train1 ditemukan *outlier* pada nilai mata kuliah dan duplikasi nilai mata kuliah. Sedangkan penelusuran terhadap data tabel data_train2 terdapat *missing value* pada kolom ipk, jumlah_sks, dan nilai.

C. Data Preparation Phase

Data preparation phase atau fase persiapan data adalah tahap memilih kasus dan parameter yang akan dianalisa, melakukan transformasi terhadap parameter tertentu, dan membersihkan data agar sesuai dengan pemodelan yang ingin dibangun dari tabel data_train1 dan data_train2.

Pengembangan *dataset* baru dengan melakukan transformasi data agar sesuai dengan kebutuhan. Adapun transformasi yang dilakukan adalah sebagai berikut :

1. Menghilangkan duplikasi nilai matakuliah dengan menyimpan hanya satu nilai terbaik dari kumpulan duplikasi setiap matakuliah yang diambil setiap mahasiswa (data_train1).
2. Mengubah *missing value* di kolom nilai menjadi "NA" (data_train1).
3. Mengubah *missing value* di kolom ipk dan jumlah_sks menjadi "NA" (data_train2).
4. Menambahkan kolom angkatan dan mengisinya dengan substring nilai kolom nim (data_train2).
5. Menambahkan kolom lulus dan mengisinya dengan substring nilai kolom tgl_pendadaran (data_train2).
6. Menghapus kode_matakuliah yang tidak termasuk dalam daftar matakuliah wajib semester I hingga semester IV (data_train1).
7. Mengelompokkan jalur PBT FK ke jalur PBT dan mengelompokkan jalur CBT luar kota ke jalur CBT.
8. Menambahkan kolom kategori sebagai kolom label atau kelas dan memberi nilai berdasarkan hasil pengurangan nilai pada kolom lulus dengan kolom angkatan. Jika hasil perhitungan adalah kurang dari sama dengan (<=) 4 maka nilai kolom kategori adalah tepat dan jika hasilnya lebih dari (>) 4 maka nilai kolom kategori adalah tidak tepat (data_train2).

Tahap integrasi dilakukan dengan menggabungkan atribut tabel data_train1 dengan tabel data_train2 *menggunakan full outer join* berdasarkan NIM yang terdapat di kedua tabel sehingga diperoleh *set* atribut terpilih meliputi NIM, ipk, jumlah_sks, kelamin, jalur, provinsi, id_matkul, nilai, angkatan, lulus, tgl_lulus, dan kategori.

Adapun proses pembersihan data adalah berikut ini :

1. Menangani *missing value* pada kolom ipk dan jumlah_sks dengan menggunakan metode *sequential method*. Konsep ini digunakan untuk atribut numerik dengan mengisi nilai yang hilang dengan rata-rata dari kelas yang sama.
2. Menangani *missing value* pada nilai dengan menggunakan nilai yang paling sering muncul atau modus pada kelas atau label yang sama.
3. Melakukan diskritisasi nilai mata kuliah dengan ketentuan berikut ini :
   - A, A- dan A/B dikelompokkan menjadi A.
   - B+, B, B- dan B/C dikelompokkan menjadi

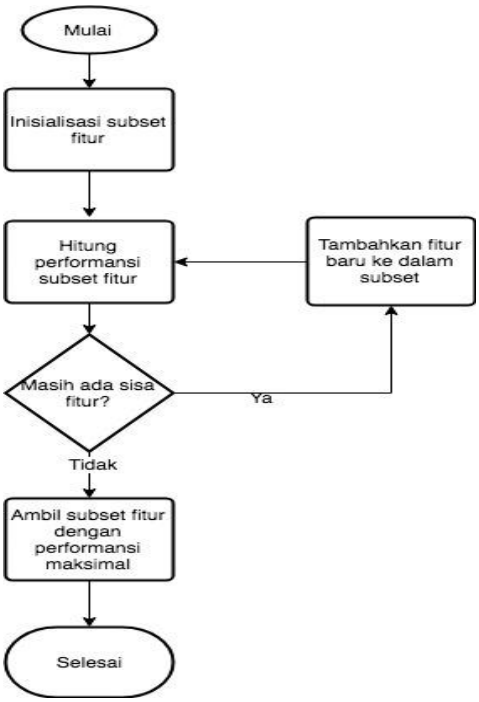
- B.
- C+, C, C-, C/D dikelompokkan menjadi C.
- D+ dan D dikelompokkan menjadi D.
- E dikelompokkan menjadi E.

4. Menghapus id_matkul yang tidak termasuk dalam daftar matakuliah wajib semester I hingga semester IV.
5. Menghapus baris data yang tidak memiliki nilai pada kolom tgl_pendadaran.

Berdasarkan dataset terbaru terdapat 692 baris data dengan 43 atribut. Atribut tersebut terdiri atas 1 atribut id, 41 atribut pemodelan, dan 1 atribut label. Agar proses klasifikasi lebih efisien dan efektif maka perlu dilakukan pemilihan fitur. Pemilihan fitur digunakan dalam pemodelan agar atribut dipilih hanya atribut yang memiliki korelasi tinggi terhadap kelulusan mahasiswa. Teknik *filter* dan *wrapper* digunakan untuk menyeleksi seluruh atribut yang ada.

Teknik *filtering* dilakukan dengan menggunakan *gain ratio attribute evaluator*, dimana setiap atribut akan dihitung nilai *gain rationya*. Jika *gain ratio* suatu atribut bernilai 0 maka atribut tersebut tidak dipilih. Kemudian dilakukan *sorting* atribut dari *gain ratio* tertinggi hingga terendah. *Sorting* atribut berdasarkan *gain ratio* ini bertujuan untuk memudahkan seleksi fitur pada tahap selanjutnya melalui teknik *wrapper*.

Pada Gambar 1 dideskripsikan alur teknik *wrapper* dalam proses pemilihan atribut. Teknik *wrapper* dilakukan dengan menghitung performansi dari setiap subset fitur melalui proses iterasi, setiap atribut ditambahkan satu per satu dalam himpunan atribut lalu diurutkan berdasarkan *score* tertinggi hingga terendah. Performansi subset fitur dihitung melalui nilai *accuracy*, *sensitivity*, dan *specificity* dari model klasifikasi dengan menggunakan algoritma pohon keputusan C4.5.



Gambar 1. Teknik *wrapping* atribut

Sensitivity didefinisikan sebagai tingkat kemampuan model untuk mengenali seluruh data lulus tepat waktu dengan benar. Sedangkan *specificity* adalah kemampuan model mengenali seluruh data lulus tidak tepat waktu dengan benar. *Accuracy* menyatakan kemampuan model dalam melakukan deteksi lulusan tepat dan tidak tepat waktu dengan benar. Ketiga metrik evaluasi ini dihitung menggunakan *confusion matrix* (1), (2), dan (3) seperti pada Tabel II berikut [13]:

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

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Nilai TP, TN, FP dan FN diperoleh dari tabel *confusion matrix.* TN adalah jumlah lulus tepat yang berhasil diprediksi sebagai lulus tepat waktu. TP adalah jumlah lulus tidak tepat yang berhasil diprediksi lulus tidak tepat. FP adalah jumlah lulus tepat waktu yang diprediksi sebaliknya. Sedangkan, FN menyatakan lulus tidak tepat waktu yang dideteksi sebaliknya.

TABEL II. CONFUSION MATRIX

Aktual	**Prediksi**	
	Tepat	***Tidak Tepat***
Tepat	TN	FP
Tidak Tepat	FN	TP

Ketika terdapat banyak kelas yang akan diambil kembali, maka membagi rata hasil pengukuran akan memberikan gambaran mengenai hasilnya secara umum. *Macro Averaging* adalah salah satu cara yang dapat digunakan untuk memperoleh rerata pengukuran evaluasi pada *multiple class.* Adapun pengukuran *precion* dan *recall* dengan *macro averaging* (4) dan (5) berikut [14]:

$$PRE_{macro} = \frac{PRE_1 + \cdots + PRE_k}{k} \quad (4)$$

Keterangan :
PRE_{macro} : *precision* atau *specificity*
PRE_1 : nilai *precision* kelas atau ke-1
PRE_k : nilai *precision* kelas atau label ke-k
K : jumlah kelas atau label

$$REC_{macro} = \frac{REC_1 + \cdots + REC_k}{k} \quad (5)$$

Keterangan:
REC_{macro} : *recall* atau *sensitivity*
REC_1 : nilai *recall* kelas atau ke-1
REC_k : nilai *recall* kelas atau label ke-k
K : jumlah kelas atau label

Untuk memilih *subset* fitur terbaik, maka diperlukan perhitungan *score* (6), (7), dan (8). Jika terdapat fitur yang memiliki *score* sama maka pilih *subset* yang memiliki nilai *gain ratio* tertinggi [13].

$$\bar{A}_i = \frac{Sens_i}{Spec_i} \quad (6)$$

$$|A|_i = |Sens_i - Spec_i| \quad (7)$$

$$Score_i = \bar{A}_i - |A|_i \quad (8)$$

Keterangan :
$Score_i$: nilai hasil evaluasi subset fitur ke –i
$\bar{A}_i$: rata-rata nilai *sensitivity* dan *specificity* subset fitur ke-i
$|A|_i$: selisih nilai *sensitivity* dan *specificity* subset fitur ke-i
$Sens_i$: nilai *sensitivity* subset fitur ke-i
$Spec_i$: nilai *specificity* subset fitur ke-i

D. Modelling Phase

Gambaran singkat proses yang dilalui saat menggali data dengan algoritma C4.5 sebagai berikut [12]:

1. Menghitung nilai *entropy* total dan *entropy* dari masing-masing atribut (9) untuk mengukur tingkat homogenitas dan *purity* dari data.

$$Entropy\ (S) = \sum_{j=1}^{k} - pj\ log_2 pj \quad (9)$$

Keterangan :
S = himpunan kasus
k = banyaknya partisi di S
Pj = probabilitas setiap partisi

2. Menghitung nilai *information gain* (10) dengan menggunakan nilai *entropy* yang telah diperoleh untuk mengukur efektifitas suatu atribut dalam mengklasifikan data.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{k} \frac{|Si|}{|S|}.Entropy\ (Si) \quad (10)$$

Keterangan :
S = himpunan kasus
A = fitur
n = jumlah partisi atribut A
|Si| = proporsi Si terhadap S
|S| = jumlah kasus dalam S

3. Menghitung nilai *split info* (11) dari setiap atribut berisi normalisasi dari *information gain* yang memperhitungkan *entropy* dari distribusi probabilitas subset setelah dilakukan proses partisi.

$$SplitInfo(S,A) = -\sum_{i=1}^{c} \frac{Si}{S} \log_2 \frac{Si}{S} \quad (11)$$

Keterangan:
S : himpunan kasus
A : atribut
Gain(S,A) :*information gain* pada atribut A
SplitInfo(S,A) = *split information* pada atribut A

4. Menghitung gain ratio (12) menggunakan nilai *information gain* dan *split info*.

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (12)$$

Keterangan :
S : himpunan kasus
A : atribut A
Gain(S,A) : *information gain* pada atribut A
SplitInfo(S,A): *split information* pada atribut A

5. Mengambil nilai *gain ratio* terbesar sebagai simpul akar.
6. Menghilangkan atribut yang sudah dipilih sebelumnya dan ulangi perhitungan nilai *entropy*, *information gain*, *split info*, dan *gain ratio* dengan memilih *gain ratio* terbesar sebagai simpul internal pohon.
7. Mengulangi perhitungan tersebut hingga semua atribut memiliki kelas.

E. Evaluation Phase

Pengujian terhadap *rules* yang terbentuk dari pemodelan data training menggunakan algoritma C45 dilakukan dengan menggunakan metode *confussion matrix*. Dengan *confussion matrix* dapat diketahui kemampuan model dalam memprediksi tingkat kelulusan mahasiswa dengan membandingkan kelulusan yang sebenarnya dan kelulusan hasil prediksi.

F. Deployment Phase

Pada fase *deployment* dilakukan implementasi model klasifikasi dalam sistem berbasis web sehingga Dosen Pembimbing Akademik dapat melakukan prediksi kelulusan mahasiswa bimbingannya.

IV. HASIL IMPELEMENTASI DAN EVALUASI

Bab ini menguraikan hasil implementasi dan evaluasi model prediksi kelulusan mahasiswa berbasis algoritma pohon keputusan.

A. Hasil Pemilihan Atribut

Adapun hasil kalkulasi subset fitur berdasakan performa *accuracy*, *sensitivy*, dan *specificity* dengan 10-*folds cross validation* diuraikan pada Tabel III. Nilai *score* tertinggi diperoleh pada kombinasi sembilan fitur yang merepresentasikan nilai mata kuliah terpilih, jumlah SKS dan IPK.

TABEL III. KALKULASI PERFORMA *SUBSET* ATRIBUT

No	Subset fitur	Akurasi	Sensitivity	Specificity	Score
1	{1}	72.11%	69.85%	73.55%	68.00 %
2	{1,2}	72.11%	69.85%	73.55%	68.00 %
3	{1,2,3}	69.51%	68.05%	69.25%	67.45 %
4	{1,2,3,4}	71.10%	69.40%	71.30%	68.45 %
5	{1,2,3,4,5}	70.81%	69.75%	70.45%	69.40 %
6	{1,2,3,4,5,6}	70.66%	69.85%	70.20%	69.68 %
7	{1,2,3,4,5,6,7}	70.66%	69.85%	70.20%	69.68 %
8	{1,2,3,4,5,6,7,8}	68.93%	68.35%	68.45%	68.30 %
9	{1,2,3,4,5,6,7,8,9}	71.10%	70.40%	70.65%	70.28 %

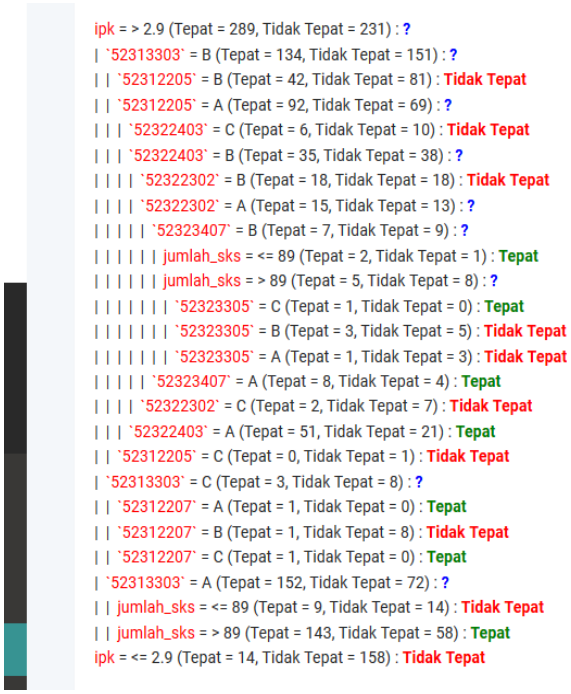
Sembilan fitur yang terpilih dari proses seleksi fitur digabungkan dengan atribut berupa identitas data dan atribut kelas (tepat dan tidak tepat). Rincian atribut yang digunakan dalam pembentukan model dijelaskan pada Tabel IV.

TABEL IV. PERHITUNGAN AKURASI DENGAN *CONFUSION MATRIX*

No	Kode Atribut	Keterangan
1	Id	Identitas data
2	IPK	Nilai indeks prestasi akumulatif
3	Jumlah_sks	Jumlah SKS yang sudah diambil
4	52322302	Nilai mata kuliah Aljabar Linier dan Matriks
5	52312207	Nilai mata kuliah Basisdata
6	52322403	Nilai mata kuliah Metode Numerik
7	52313303	Nilai mata kuliah Sistem Informasi
8	52323407	Nilai mata kuliah Pemrograman Web
9	52323305	Nilai mata kuliah Pemrograman Berorientasi Obyek
10	52312205	Nilai mata kuliah Sistem Operasi
11	Kategori	Atribut

B. *Hasil Pemodelan*

Dari tahap sebelumnya diperoleh bahwa *data training* terdiri dari 692 baris data dan 11 kolom. Dataset tersebut akan digunakan untuk pembentukan pohon keputusan, pemangkasan cabang pohon, dan evaluasi akurasi *rules* yang diperoleh. Selanjutnya dilakukan pemangkasan cabang dengan *post*



Gambar 2. Aturan klasifikasi

pruning error estimate menggunakan *confident level* sebesar 25% atau z-score 0.69. Nilai ini diperoleh berdasarkan hasil pengujian terhadap beberapa nilai *confident level* yang menunjukkan bahwa nilai akurasi tertinggi *confusion matrix* dicapai saat *confident level* bernilai 25%. Pemangkasan dilakukan terhadap 100% data training sehingga diperoleh *rules* pohon keputusan seperti

Gambar 2.

C. *Hasil Evaluasi*

Pengujian *confussion matrix* dilakukan dengan membandingkan kelulusan yang sebenarnya dan kelulusan hasil prediksi. Adapun hasil pengujian yang diperoleh dari Tabel vV.

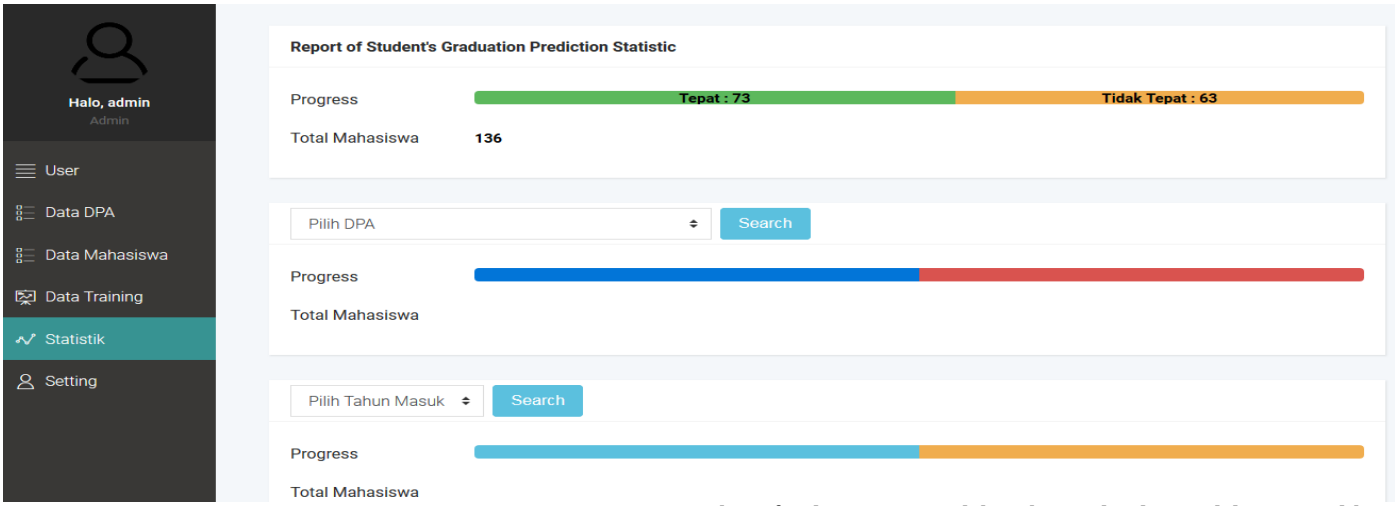
TABEL V. PERHITUNGAN AKURASI DENGAN *CONFUSION MATRIX*

Aktual	Prediksi	
	Tepat	***Tidak Tepat***
Tepat	207	96
Tidak Tepat	84	305

- *Accuracy :* 73.99%, bahwa kemampuan deteksi sistem dalam melakukan deteksi lulusan tepat dan tidak tepat dengan benar adalah sebesar 73.99%.
- *Specificity :* 78.41 %, bahwa tingkat kemampuan metode mengenali seluruh data lulus tidak tepat dengan benar adalah sebesar 78.41%.
- *Sensitivity :* 68.32 %, bahwa tingkat kemampuan metode untuk mengenali seluruh data lulus tepat dengan benar adalah sebesar 68.32%.

Dari hasil di atas dapat dilihat bahwa performansi model belum terlalu optimal dalam melakukan prediksi. Salah satu

penyebabnya adalah adanya ketidakseimbangan data di mana jumlah data histori mahasiswa yang lulus tidak tepat waktu lebih banyak dibandingkan data histori mahasiswa yang lulus tepat waktu. Oleh karena itu, dapat dipertimbangkan untuk mengambil data mahasiswa di angkatan 2014 yang telah lulus secara tepat waktu.



Gambar 3. Antarmuka laporan prediksi

D. Hasil Deployment

Model klasifikasi yang sudah terbentuk berupa aturan pohon keputusan diaplikasikan dalam sebuah sistem berbasis *web* dengan pengguna terdiri dari Admin, mahasiswa dan Dosen Pembimbing Akademik. Pada **Error! Reference source not found.** disajikan fitur Laporan prediksi ketepatan masa studi mahasiswa Teknik Informatika UII. Data yang ingin ditampilkan dapat dipilih berdasarkan DPA dan Tahun Masuk Mahasiswa.

V. SIMPULAN DAN SARAN

Berdasarkan hasil analisa, implementasi, dan pengujian terhadap penelitian ini maka diperoleh kesimpulan bahwa algoritma C.45 dapat digunakan untuk memprediksi ketepatan masa studi mahasiswa dengan menggunakan data latih mahasiswa Teknik Informatika UII angkatan 2010 hingga 2013. Dengan menggunakan teknik *filter dan wrapper* terhadap subset atribut, maka ditemukan pengetahuan bahwa IPK, jumlah SKS, nilai matakuliah Aljabar Linear dan Matriks, Basisdata, Metode Numerik, Sistem Informasi, Pemrograman Web, Pemrograman Berorientasi Obyek, dan Sistem Operasi memiliki pengaruh yang cukup signifikan terhadap tingkat kelulusan mahasiswa.

Pengujian terhadap model klasifikasi berbasis pohon keputusan C4.5 memberikan hasil yang cukup baik, yaitu dengan dicapainya akurasi sebesar 73.9%. Pada fase *deployment*, dilakukan penerapan *rules* ke dalam aplikasi berbasis *web* yang dapat digunakan oleh mahasiswa dan Dosen Pembimbing Akademik. Selain itu, sistem yang dikembangkan dapat menangani pembentukan *rules* yang dinamis.

Untuk pengembangan selanjutnya, diharapkan model dapat digunakan untuk memprediksi kelulusan mahasiswa Jurusan Teknik Informatika UII dengan pengaturan kurikulum yang dinamis. Selain itu, dapat dilakukan penerapan algoritma klasifikasi dengan karakter yang berbeda dengan pohon keputusan seperti SVM dan Naïve bayes. Penerapan SVM diharapkan dapat memberikan model generalisasi yang lebih baik dengan terlebih dahulu menganalisis data latihnya apakah bersifat linier atau tidak sebagai landasan dalam pemilihan kernel. Penerapan Naïve Bayes diharapkan dapat memperbaiki performansi karena kemampuannya dalam menangani kondisi kuantitas data latih yang tidak terlalu representatif.

REFERENSI

[1] M. M. Quadri, N. V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," *Global Journal of Computer Science and Technology*, 2010.

[2] W. Xindong et al, "Top 10 Algorithms in Data Mining," *Proceeding of IEEE Conference on Data Mining*, 2009.

[3] P. Chapman et al, "CRISP-DM 1.0," *CRISP-DM Consorsium*, 2000. [Online] Available: dari https://www.the-modeling-agency.com/crisp-dm.pdf

[4] Shahiri et al, "A Review of Predicting Student's Performance Using Data Mining Techniques," *Elsevier Procedia Computer Science*, vol. 72, pp 412-422, 2015.

[5] S. M. S. Damanik et al, "Klasifikasi Lama Studi Mahaiswa FSM Universitas Diponegoro Menggunakan Regresi Logistik Biner dan Support Vector Machine (SVM)," *Jurnal Gaussian*, vol. 4, no. 1, 2015, 123-132.

[6] Suniantara, K. Putu, and M. Rusli, "Klasifikasi Waktu Kelulusan Mahasiswa STIKOM Bali Menggunakan CHAID Regression - Trees dan Regresi Logistik Biner," *Jurnal Statistika Universitas Muhammadiyah Semarang*, vol. 5, no. 1, 2017, pp. 27-32.

[7] Y. K. Sari, "Implementasi Klasifikasi Data Mining untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes," *Generation Journal Universitas Amikom Yogyakarta* , 2017, pp. 96-103

[8] Y. S. Nugroho, Y. S., "Penerapan Algoritma C4.5 untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta," *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, 2014, pp. A 1-6.

[9] Prasetyo et al, "Prediksi Kelulusan Mahasiswa Pada Perguruan Tinggi Kabupaten Majalengka Berbasis Knowledge Based System," *Prosiding Seminar Nasional Telekomunikasi dan Informatika*, 2016.

[10] S. Gunawan, "Pembentukan Model Klasifikasi Data Lama Studi Mahasiswa Stmik Indonesia Menggunakan Decision Tree Dengan Algoritma Nbtree," *Prosiding Seminar Nasional Teknologi Informasi dan Aplikasi Komputer*, 2017.

[11] A. Rohman, "Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa," *Jurnal Neo Teknika*, vol. 1, no. 1, 2015.

[12] Larose, D. T., *Discovering Knowledge in Data : An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc, 2015.

[13] I. Z. Mutaqien, *"Pengembangan Metode Seleksi Fitur dan Transformasi Data pada Sistem Deteksi Intrusi dengan Pembatasan Ukuran Cluster dan Sub-Medoid,"* Doctoral dissertation, Institut Teknologi Sepuluh Nopember, 2017.

[14] S. Raschka, "Machine Learning FAQ," *Sebastian Raschka*, January 2018. [Online]. Available: Sebastian Raschka, https://sebastianraschka.com/faq/docs/multiclass-metric.html