khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika

Ratna Puspita Sari Putri*, Indra Waspada
Departemen Ilmu Komputer/Informatika, Fakultas Sains dan Matematika
Universitas Diponegoro
Semarang
*ratnapsp@gmail.com

Abstrak-Data tentang mahasiswa yang lulus merupakan sebuah data yang penting baik bagi departemen, fakultas maupun universitas karena data tersebut digunakan dalam proses akreditasi. Data tentang mahasiswa yang lulus terus bertambah di tiap tahunnya dan menumpuk seperti data yang terabaikan karena jarang digunakan. Data tentang mahasiswa yang lulus dapat memberikan informasi yang berguna jika dimanfaatkan dengan maksimal. Maka dari itu, penelitian ini akan memanfaatkan data tentang mahasiswa yang lulus dengan mengolahnya menggunakan *data mining* untuk mendapatkan informasi berupa prediksi kelulusan mahasiswa. Metode yang akan digunakan adalah metode pohon keputusan yang dibangun dengan algoritma C4.5 disertai dengan algoritma *error-based pruning* untuk proses pemotongan pohon keputusan. Kriteria yang akan digunakan adalah jenis kelamin, asal daerah, IPK, dan TOEFL. Dalam penerapannya, algoritma C4.5 dapat digunakan untuk menghasilkan prediksi kelulusan dengan nilai rata-rata *precision* 63.93%, *recall* 60.73%, dan akurasi **60.52%**. Setelah pohon keputusan dipotong dengan menggunakan metode *error-based pruning*, didapatkan hasil yang lebih baik. Pohon yang dipotong dengan menggunakan nilai *confidence* 0,4 menghasilkan *precision* 70.70%, *recall* 50.65%, dan akurasi 61.57%. Sedangkan pohon yang dipotong dengan menggunakan nilai *confidence* 0,25 menghasilkan *precision* 73.77%, *recall* 48.84%, dan akurasi 62.44%.

Kata Kunci: *data mining*; kelulusan mahasiswa, pohon keputusan, C4.5, *error-based pruning*

1. Pendahuluan

Di era digital ini banyak instansi dan perusahaan yang telah menyimpan data mereka di dalam sebuah *database* yang terkomputerisasi. Dunia pendidikan pun tidak terlepas dari perkembangan teknologi ini. Universitas Diponegoro termasuk salah satu perguruan tinggi yang telah menyimpan datanya dalam *database* yang terkomputerisasi. Data tersebut merupakan data mahasiswa, data dosen, serta berbagai data lain yang berhubungan dengan Universitas Diponegoro.

Data tersebut tidak banyak memiliki kegunaan dan seolah-olah menjadi sekumpulan data terabaikan yang bertambah besar tiap tahunnya. Data tersebut hanya digunakan saat universitas membutuhkan suatu informasi tertentu atau saat proses akreditasi. Saat mahasiswa telah lulus maka data mereka akan semakin jarang digunakan. Padahal data tentang mahasiswa yang lulus merupakan data yang penting dan digunakan dalam proses akreditasi.

Data tentang mahasiswa yang lulus dapat memberikan informasi yang berguna bagi universitas jika dimanfaatkan dengan maksimal. Salah satu cara untuk memanfaatkan data tentang mahasiswa yang lulus ini adalah dengan mengolahnya menggunakan *data mining*. Dengan proses *data mining* ini dapat ditemukan pola atau aturan yang dapat digunakan untuk menghasilkan suatu informasi seperti prediksi kelulusan mahasiswa.

Prediksi kelulusan mahasiswa dapat digunakan lebih lanjut untuk membantu universitas dalam mengevaluasi dan memperbaiki sistem pembelajaran sehingga universitas dapat menghasilkan lulusan yang berkualitas.

Penelitian ini akan dilaksanakan di Prodi Informatika. Prodi Informatika telah berdiri dari tahun 2004 dan memiliki sasaran untuk menjadi program studi unggulan. Oleh karena itu hasil prediksi kelulusan mahasiswa dapat membantu Prodi Informatika dalam mengambil langkah strategis.

Dalam penelitian yang berjudul *Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques*, Jadhav dan Channe membandingkan performa metode K-NN, Naïve Bayes, dan pohon keputusan dalam berbagai aspek dengan menggunakan berbagai *dataset*. Dari hasil penelitian tersebut dapat diketahui bahwa pohon keputusan merupakan metode yang paling cepat performanya dibandingkan dengan metode yang lain. Selain itu pohon keputusan lebih akurat dan memiliki *error rate* yang rendah [1].

Dalam penelitian lain yang berjudul *Comparative Analysis of Decision Tree Algorithms for The Prediction of Eligibility of A Man for Availing Bank Loan*, Mohankumar dkk membandingkan berbagai algoritma untuk membangun pohon keputusan dan algoritma C4.5 merupakan algoritma dengan performa tercepat dan memiliki akurasi yang paling tinggi [2].

David Kamagi mengimplementasikan algoritma C4.5 dalam penelitian berjudul "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa" dan menghasilkan prediksi dengan tingkat keakuratan yang tinggi, yaitu 87,5% [3]. Kamagi menggunakan empat kelas target, yaitu lulus cepat, lulus tepat, lulus terlambat, dan *drop out*. Atribut yang digunakan adalah IPS, jenis kelamin, asal sekolah, tipe kelulusan, dan

jumlah SKS. Dalam penelitian ini akan digunakan kelas target <5 tahun dan ≥5 tahun. Sedangkan atribut yang digunakan adalah IPK, TOEFL, asal daerah, dan jenis kelamin.

Berdasarkan penjelasan di atas akan dibangun sebuah aplikasi prediksi kelulusan mahasiswa Prodi Informatika dengan menggunakan metode pohon keputusan yang dibangun menggunakan algoritma C4.5.

2. Metode

a. Pemahaman Domain dan Tujuan KDD

Data tentang mahasiswa yang lulus merupakan sebuah data yang penting karena data tersebut digunakan dalam proses akreditasi. Dalam penelitian sebelumnya telah dibangun sebuah aplikasi repositori lulusan yang digunakan untuk menyimpan dan mengelola data tentang mahasiswa Prodi Informatika yang lulus. Setelah disimpan data tersebut tidak terlalu banyak digunakan dan semakin bertambah besar tiap tahunnya.

Data tentang mahasiswa yang lulus dapat memberikan informasi yang berguna bagi prodi jika dimanfaatkan dengan maksimal. Salah satu cara untuk memanfaatkan data tentang mahasiswa yang lulus ini adalah dengan mengolahnya menggunakan *data mining*. Dengan proses *data mining* ini dapat ditemukan pola atau aturan yang dapat digunakan untuk menghasilkan suatu informasi seperti prediksi tepat atau tidaknya kelulusan mahasiswa.

Untuk itu akan dilakukan proses penggalian informasi dari data tentang mahasiswa yang lulus Prodi Informatika dengan menggunakan model *knowledge discovery in databases* (KDD). Tujuan yang diharapkan dari proses KDD ini adalah mendapatkan informasi mengenai kelulusan seorang mahasiswa berdasarkan jenis kelamin, asal daerah, IPK, serta nilai TOEFL-nya.

b. Pemilihan dan Penambahan Data

Data yang akan digunakan dalam proses KDD adalah data tentang mahasiswa yang lulus Prodi Informatika dari Januari 2013 sampai Agustus 2017. Data tentang mahasiswa yang lulus tersebut diambil dari aplikasi repositori lulusan. Tidak semua data tentang mahasiswa yang lulus akan digunakan. Data yang akan digunakan adalah nama mahasiswa, jenis kelamin, asal daerah, IPK, nilai TOEFL, dan lama studi. Jenis kelamin, asal daerah, IPK, dan nilai TOEFL akan digunakan sebagai atribut dan lama studi akan digunakan sebagai kelas.

Dalam pembangunan aplikasi prediksi kelulusan mahasiswa, IPK menggambarkan performa akademik mahasiswa. Nilai TOEFL menggambarkan pemahaman mahasiswa dalam memahami literatur pembelajaran yang menggunakan bahasa Inggris. Asal daerah menggambarkan pengaruh faktor keluarga dan perbedaan kultur terhadap performa akademik mahasiswa. Keluarga dan kemandirian belajar merupakan faktor yang menentukan prestasi mahasiswa [4]. Selain itu fenomena *culture shock* seringkali terjadi pada mahasiswa perantauan. *Culture shock* tersebut dapat menimbulkan efek stres yang dapat mempengaruhi prestasi mahasiswa [5]. Sedangkan jenis kelamin menggambar pengaruh gender terhadap performa akademik mahasiswa. Dalam sebuah penelitian yang meneliti pengaruh gender dan motivasi belajar terhadap prestasi siswa, perempuan dinilai lebih berprestasi daripada laki-laki. Hal ini dikarenakan perempuan lebih tekun dan rajin daripada laki-laki [6].

Tabel 1. Tabel Transformasi Data

No.	Atribut	Nilai Atribut	Keterangan
1.	Asal daerah	Jateng	Provinsi bernilai Jawa Tengah
		Luar Jateng	Provinsi selain Jawa Tengah yang masih berada di Pulau Jawa
		Luar Jawa	Provinsi di luar Pulau Jawa
2.	IPK	Memuaskan	IPK kurang dari 2,75
		Sangat Memuaskan	IPK di antara 2,76 sampai dengan 3,50
		Dengan Pujian	IPK lebih dari 3,50
3.	TOEFL	Dasar	Nilai TOEFL kurang dari 420
		Menengah Bawah	Nilai TOEFL di antara 421 sampai 480
		Menengah Atas	Nilai TOEFL di antara 481 sampai 520
		Mahir	Nilai TOEFL lebih dari 520
4.	Lama studi	<5 tahun	lama studi kurang dari 5 tahun
		≥5 tahun	lama studi lebih dari sama dengan 5 tahun

Jumlah data yang digunakan adalah 382 data dengan 212 data untuk kelas <5 tahun dan 170 data untuk kelas ≥5 tahun.

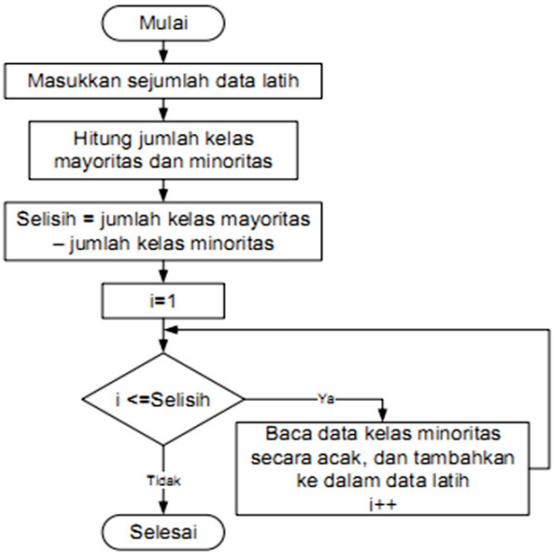
c. Pembersihan dan Pemrosesan Awal Data

Pembersihan data dilakukan untuk membersihkan *noise* pada data. Dalam penelitian ini proses pembersihan data dilakukan dalam aplikasi repositori lulusan. Aplikasi repositori lulusan telah diatur sedemikian rupa sehingga setiap data tentang mahasiswa yang lulus termasuk data nama, jenis kelamin, asal daerah, IPK, nilai TOEFL, dan lama studi yang disimpan tidak kosong dan konsisten.

d. Transformasi Data

Transformasi data dilakukan untuk mengubah data menjadi nilai dengan format tertentu. Dalam proses KDD ini akan digunakan data yang bersifat diskrit, oleh karena itu data yang bersifat kontinu akan diubah menjadi data diskrit. Selain itu ada data yang cakupannya terlalu luas dan akan mempengaruhi proses KDD sehingga perlu dikelompokkan menjadi beberapa kelompok kecil.

Data asal daerah merupakan data yang cakupannya luas. Asal daerah dalam data lulusan berisi nama provinsi asal dari masing-masing mahasiswa. Asal daerah akan dibagi menjadi 3 (tiga) kelompok, yaitu Jateng (Jawa Tengah), luar Jateng (luar Jawa Tengah), dan luar Jawa.



Gambar 1. *Flowchart* ROS

Pengelompokan IPK akan dibagi menjadi 3 kategori, yaitu tinggi, sangat memuaskan, dan rendah. Nilai per kelompok dibagi berdasarkan Peraturan Rektor Universitas Diponegoro Tahun 2012 Pasal 20. Untuk mahasiswa dengan IPK kurang dari 2,75 akan dikelompokkan sebagai IPK memuaskan, mahasiswa dengan IPK antara 2,76 sampai 3,50 akan dikelompokkan sebagai IPK sangat memuaskan, dan mahasiswa dengan IPK lebih dari 3,50 akan dikelompokkan sebagai IPK dengan pujian.

Dalam penelitian yang berjudul "*Reading-Writing Relationship in First and Second Language*", Carson dkk mengelompokkan nilai TOEFL menjadi 4 kelas, yaitu dasar, menengah bawah, menengah atas, dan mahir. Nilai TOEFL kurang dari 420 termasuk dalam kelas dasar, nilai di antara 421 sampai 480 termasuk dalam kelas menengah bawah, nilai di antara 481 sampai 520 termasuk dalam kelas menengah atas, dan untuk nilai di atas 520 termasuk dalam kelas mahir [7].

Pada penelitian ini kelulusan seorang mahasiswa S1 Universitas Diponegoro akan dibagi menjadi dua kelompok, yaitu <5 tahun dan ≥5 tahun.

e. Data Mining

Untuk mencapai tujuan dari proses KDD akan digunakan metode pohon keputusan sebagai metode *data mining* dan algoritma C4.5 untuk membangun pohon keputusan. Untuk proses pemotongan pohon keputusan akan digunakan metode *error-based pruning*. Selain itu dikarenakan adanya *imbalance data*, data yang akan digunakan dalam proses pembuatan pohon keputusan akan ditangani dengan menggunakan algoritma *random over sampling* (ROS).

Tahap *data mining* dimulai dengan membagi data tentang mahasiswa yang lulus yang telah ditransformasi menjadi data latih dan data uji. Data latih akan digunakan dalam proses pembuatan pohon keputusan. Sedangkan data uji akan digunakan untuk mengukur kinerja dari pohon keputusan yang telah dibuat. Data latih yang akan digunakan dalam proses pembuatan pohon keputusan harus diseimbangkan terlebih dahulu untuk menghindari adanya kecenderungan terhadap kelas mayoritas dalam pohon yang dibuat. Proses untuk menyeimbangkan data latih dilakukan dengan menggunakan metode *random over sampling* (ROS). Metode ROS dilakukan dengan menghitung selisih dari kelas mayoritas dan kelas minoritas. Kemudian dipilih satu data secara acak dari kelas minoritas. Data tersebut lalu ditambahkan ke dalam kelas minoritas. Penambahan data akan diulangi sampai jumlah data dalam kelas mayoritas dan kelas minoritas sama [8]. Metode ROS digambarkan dengan *flowchart* pada gambar 1.

Setelah data latih telah seimbang, data latih siap digunakan dalam proses membuat pohon keputusan dengan menggunakan algoritma C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang ditemukan oleh J. Ross Quinlan pada tahun 1993 [9]. Algoritma C4.5 adalah sebagai berikut [10]:

1) Menghitung jumlah kasus total, jumlah kasus dengan keputusan <5 tahun, kasus dengan keputusan ≥5 tahun, dan *entropy* dari semua kasus dan kasus yang dibagi berdasarkan nilai atribut.

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

S = kumpulan data
k = banyaknya kelas dalam S
p_j = probabilitas kelas C_j

Jika kasus total hanya memiliki satu kelas (<5 tahun atau ≥5 tahun), maka jadikan node sebagai daun dengan nilai kelas yang mayoritas.

2) Menghitung *information gain* untuk setiap atribut.

$$gain(A) = entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} x\ entropy(S_i) \quad (2)$$

S = kumpulan data
A = atribut
A_i = nilai atribut ke-i
$|S_i|$ = jumlah data untuk A_i
$|S|$ = jumlah data dalam S
k = jumlah nilai atribut

3) Menghitung *split info* untuk setiap atribut.

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (3)$$

S = kumpulan data
A = atribut
A_i = nilai atribut ke-i
n = jumlah nilai atribut
S_i = jumlah data untuk A_i

4) Menghitung *gain ratio* untuk setiap atribut.

$$GainRatio(A) = \frac{gain(A)}{SplitInfo(S,A)} \quad (4)$$

A = atribut
S = kumpulan data

5) Memilih atribut dengan nilai *gain ratio* terbesar sebagai *node*.
6) Membagi data berdasarkan nilai atribut dari atribut terpilih. Kemudian menggunakannya untuk melakukan langkah selanjutnya.
7) Ulangi langkah 1 sampai 6 hingga seluruh atribut digunakan atau memenuhi suatu kondisi berhenti.

Untuk pemotongan pohon keputusan akan digunakan metode *error-based pruning*. *Error-based pruning* seringkali dideskripsikan sebagai *pessimistic error pruning* yang

ditambah dengan kemungkinan untuk mengganti *parent node* dengan *child* bernilai maksimum. Namun sebenarnya perhitungan estimasi *pessimistic error* dalam dua metode tersebut dilakukan dengan cara yang sama sekali berbeda [9]. Dalam metode *Error-Based Pruning* pemotongan pohon dilakukan dengan cara menilai setiap *node* bukan daun dari bawah pohon. Jika pengganti sub-pohon dengan daun akan membuat estimasi *error rate* lebih rendah, maka pohon akan dipotong. Setiap estimasi *error rate* untuk semua pohon yang termasuk di dalam sub-pohon ini akan terpengaruh. Karena *error rate* untuk keseluruhan pohon akan menurun seiring dengan menurunnya *error rate* dari sub-pohon, proses ini akan membuat sebuah pohon keputusan dengan *error rate* minimal, sehubungan dengan cara pemotongan pohon yang dilakukan [10]. Metode *error-based pruning* adalah sebagai berikut:

1) Menghitung jumlah kasus total, jumlah kasus dengan keputusan <5 tahun, dan kasus dengan keputusan ≥5 tahun dari sub pohon yang akan dihitung.
2) Menghitung nilai f untuk masing-masing *node* pada sub pohon. Nilai didapatkan dengan membagi jumlah kasus dengan keputusan kelas minoritas dengan jumlah kasus total.
3) Menghitung *error estimate* dari setiap *node* pada sub pohon.

$$e = \frac{f+\frac{z^2}{2N}+z\sqrt{\frac{f(1-f)}{N}+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}} \quad (5)$$

e = estimasi *error rate*
f = kesalahan dalam data pelatihan
N = jumlah kasus pada *node*
z = *confidence limit,* dimana $z = z_1 - (\alpha/2)$ untuk *confidence level* α

4) Menghitung *error estimate* rata-rata untuk *node* anak sesuai dengan rasionya. Rasio untuk *node* anak dihitung dengan membagi jumlah kasus *node* anak dengan jumlah kasus node orang tua.
5) Membandingkan *error estimate* anak dan orang tua. Jika *error estimate* orang tua lebih kecil dari anak, maka pohon akan dipotong. Lalu *node* orang tua akan diubah menjadi daun dengan nilai kelas mayoritas. Sebaliknya jika node anak lebih kecil dari orang tua, pohon tidak dipotong.
6) Ulangi langkah 1 sampai 5 hingga seluruh sub pohon diperiksa.

3. Hasil

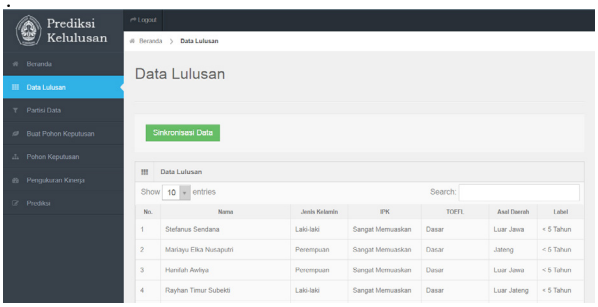
Pada bagian ini disajikan hasil penelitian yang berupa implementasi aplikasi prediksi kelulusan serta interpretasi dan evaluasinya.

a. Implementasi

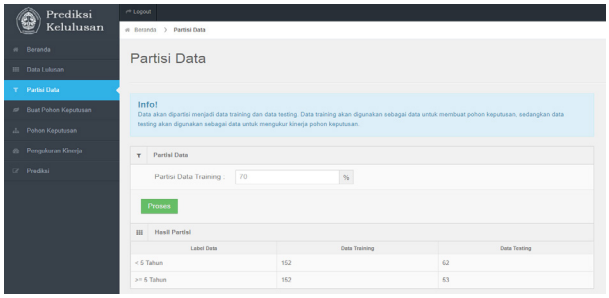
Implementasi program dilakukan berdasarkan kebutuhan fungsional yang tertera pada tabel 2. Untuk implementasi dari fungsional menampilkan data lulusan (SRS-PrediksiMhs-F-01) dapat dlihat pada gambar 2, berikutnya adalah implementasi dari fungsional melakukan dan menampilkan partisi data (SRS-PrediksiMhs-F-02) untuk memisahkan antara data *training* dengan data uji, disajikan pada gambar 3

Tabel 2. Kebutuhan fungsional

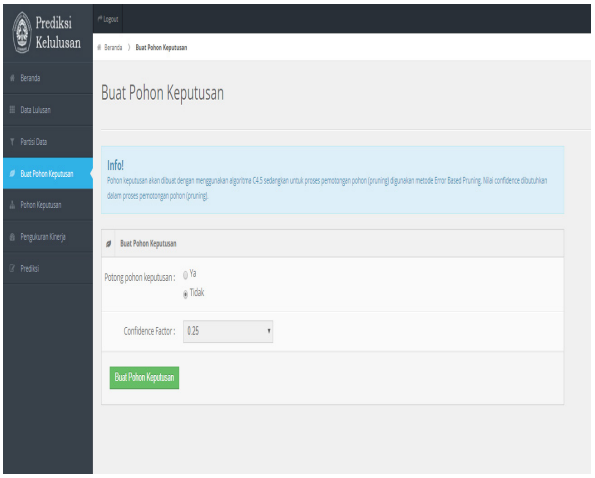
No.	SRS-ID	Deskripsi
1.	SRS-PrediksiMhs-F-01	Menampilkan data lulusan
2.	SRS-PrediksiMhs-F-02	Melakukan dan menampilkan partisi data
3.	SRS-PrediksiMhs-F-03	Membuat dan menampilkan pohon keputusan
4.	SRS-PrediksiMhs-F-04	Melakukan dan menampilkan hasil pengukuran kinerja
5.	SRS-PrediksiMhs-F-05	Mengidentifikasi kelulusan mahasiswa berdasarkan kriteria yang ditentukan



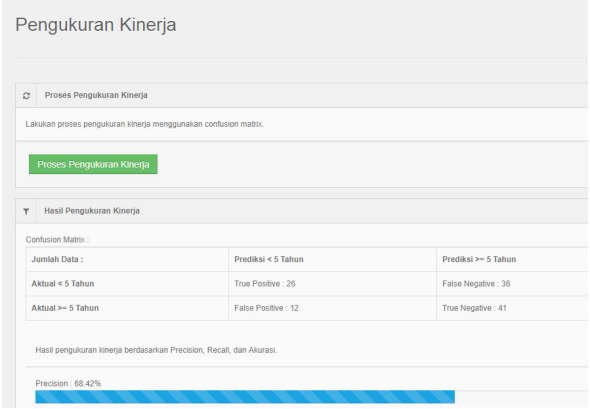
Gambar 2. Implementasi fungsional menampilkan data lulusan



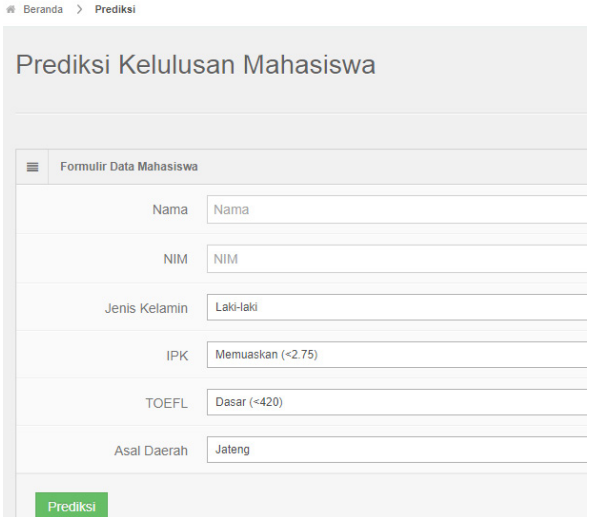
Gambar 3. Implementasi fungsional melakukan dan menampilkan partisi data



Gambar 4. Implementasi fungsional membuat dan menampilkan pohon keputusan



Gambar 5. Implementasi fungsional melakukan dan menampilkan hasil pengukuran kinerja



Gambar 6. Implementasi fungsional mengidentifikasi kelulusan mahasiswa berdasarkan kriteria yang ditentukan

Gambar 7. Hasil prediksi kelulusan mahasiswa

Implementasi dari fungsional membuat dan menampilkan pohon keputusan (SRS-PrediksiMhs-F-03) beserta pengaturan nilai *confidence factor* dapat dlihat pada gambar 4. Implementasi dari fungsional melakukan pengukuran kinerja dan menampilkan hasilnya (SRS-PrediksiMhs-F-04) disajikan pada gambar 5.

Yang terakhir adalah fitur mengidentifikasi kelulusan mahasiswa berdasarkan kriteria yang ditentukan (SRS-PrediksiMhs-F-05) dapat dlihat tampilan implementasinya pada gambar 6, yang hasilnya disajikan pada gambar 7.

Tabel 3. Detail pengukuran kinerja

Pengukuran ke-	Data Latih	Data Uji	*Pruning*	Nilai *confidence*
1	304	115	Tidak	-
2	304	115	Ya	0,25
3	304	115	Ya	0,4
4	308	115	Tidak	-
5	308	115	Ya	0,25
6	308	115	Ya	0,4
7	302	115	Tidak	-
8	302	115	Ya	0,25
9	302	115	Ya	0,4
10	298	115	Tidak	-
11	298	115	Ya	0,25
12	298	115	Ya	0,4
13	290	115	Tidak	-
14	290	115	Ya	0,25
15	290	115	Ya	0,4

Tabel 4. Confusion matrix pengukuran ke-1

	Prediksi <5 tahun	Prediksi ≥5 tahun
Aktual <5 tahun	*True positives:* 38	*False negatives:* 22
Aktual ≥5 tahun	*False positives:* 23	*True negatives:* 32

Tabel 5. Hasil pengukuran kinerja

Pengukuran ke-	*Precision*	*Recall*	Akurasi
1	62.3%	63.33%	60.86%
2	69.7%	38.33%	59.13%
3	69.7%	38.33%	59.13%
4	64%	55.17%	61.74%
5	66.07%	63.79%	65.22%
6	64.58%	53.45%	61.74%
7	60.66%	60.66%	58.26%
8	69.23%	59.02%	64.35%
9	69.23%	59.02%	64.35%
10	65.52%	60.32%	60.87%
11	83.87%	41.27%	63.48%
12	83.87%	41.27%	63.48%
13	67.19%	64.18%	60.87%
14	80%	41.79%	60%
15	66.13%	61.19%	59.13%

Tabel 6. Hasil rata-rata pengukuran kinerja

Pruning	Nilai *confidence*	*Precision*	*Recall*	Akurasi
Tidak	-	63.93%	60.73%	60.52%
Ya	0,25	73.77%	48.84%	62.44%
Ya	0,4	70.70%	50.65%	61.57%

b. Interpretasi

Pola yang dihasilkan dari proses *data mining* dapat ditampilkan dalam bentuk pohon keputusan dan *rules*. Berikut adalah contoh pohon keputusan yang dihasilkan:

```
IPK = Sangat Memuaskan = ≥5 Tahun
IPK = Dengan Pujian
| TOEFL = Dasar = <5 Tahun
| TOEFL = Menengah Bawah = <5 Tahun
| TOEFL = Menengah Atas = <5 Tahun
| TOEFL = Mahir = ≥5 Tahun
IPK = Memuaskan = ≥5 Tahun
```

Dari pohon keputusan tersebut dapat diambil *rules* sebagai berikut:

1. IF IPK sangat memuaskan THEN ≥5 tahun
2. IF IPK dengan pujian AND TOEFL dasar THEN <5 tahun
3. IF IPK dengan pujian AND TOEFL menengah bawah THEN <5 tahun
4. IF IPK dengan pujian AND TOEFL menengah atas THEN <5 tahun
5. IF IPK dengan pujian AND TOEFL mahir AND asal daerah Jateng AND jenis kelamin perempuan THEN <5 tahun
6. IF IPK dengan pujian AND TOEFL mahir THEN ≥5 tahun
7. IF IPK memuaskan THEN ≥5 tahun

c. Pengukuran Kinerja Pohon Keputusan

Pengukuran kinerja dilakukan untuk mengevaluasi pohon keputusan yang telah dibangun sebelumnya. Pengukuran kinerja dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* didapatkan dengan melakukan perbandingan hasil prediksi dari aplikasi dan hasil sebenarnya. Untuk mengetahui kinerja dari pohon keputusan akan digunakan *precision, recall,* dan akurasi.

Proses pengukuran kinerja akan dilakukan sebanyak 15 kali terhadap 5 data uji berbeda. Lima pengukuran kinerja akan dilakukan terhadap pohon keputusan tanpa pemotongan dan sisanya dilakukan terhadap pohon keputusan yang telah dipotong. Untuk pohon keputusan dengan pemotongan, proses pemotongan pohon dilakukan dengan nilai *confidence* sebesar 0,25 dan 0,4. Sedangkan data uji yang digunakan dalam proses pengukuran kinerja adalah sebesar 30% dari data tentang mahasiswa yang lulus yang telah di transformasi. Detail dari pengukuran kinerja yang akan dilakukan dapat dilihat pada tabel 3.

Untuk melakukan pengukuran kinerja pohon keputusan, langkah pertama yaitu hitung *true positives, true negatives, false positives,* dan *false negatives* dari setiap pengukuran, kemudian masukkan hasilnya ke dalam *confusion matrix*.

Langkah kedua, hitung *precision, recall, dan* akurasi. Akurasi didefinisikan sebagai tingkat kedekatan antara nilai hasil prediksi dengan nilai aktual. *Precision* didefinisikan sebagai pengukuran ketepatan. Jika data diprediksi positif, seberapa seringkah data prediksi tersebut benar. Sedangkan *recall* didefinisikan sebagai pengukuran kelengkapan. Dari jumlah data sebenarnya yang bernilai positif, sebanyak apakah data yang diprediksi positif [11].

$$Precision = TP / (TP+FP) \quad (6)$$

$$Recall = TP / (TP+FN) \quad (7)$$

$$Akurasi = (TP+TN) / N \quad (8)$$

TP = nilai *true positives*
TN = nilai *true negatives*
FP = nilai *false positives*
FN = nilai *false negatives*
N = jumlah data

Kemudian ubah nilai *precision, recall,* dan akurasi menjadi nilai persentase. Hasil dari pengukuran yang telah dilakukan dapat dilihat pada tabel 5.

Hasil pada tabel 5 kemudian dirata-rata berdasarkan pohon keputusan tanpa pemotongan, pohon keputusan yang dipotong dengan nilai *confidence* 0,25, dan pohon keputusan yang dipotong dengan nilai *confidence* 0,4.

Pengukuran kinerja yang dilakukan terhadap pohon keputusan tanpa pemotongan menghasilkan nilai *precision* sebesar 63,93%, nilai *recall* sebesar 60,73%, dan nilai akurasi sebesar 60,52%. Sedangkan untuk pohon keputusan yang dipotong dengan nilai *confidence* 0,25 menghasilkan nilai *precision* sebesar 73,77%, nilai *recall* sebesar 48,84%, dan nilai akurasi sebesar 62,44%. Pengukuran kinerja untuk pohon keputusan yang dipotong dengan nilai *confidence* 0,4 menghasilkan nilai *precision* sebesar 70,70%, nilai *recall* sebesar 50,65%, dan nilai akurasi sebesar 61,57%. Dari hasil tersebut dapat diketahui bahwa pemotongan pohon keputusan dengan menggunakan metode *error-based pruning* dapat meningkatkan akurasi. Pemotongan dengan menggunakan nilai *confidence* sebesar 0,25 meningkatkan akurasi lebih baik daripada nilai *confidence* 0,4.

d. Mengolah Pengetahuan

Hasil dari proses KDD dalam penelitian ini adalah prediksi apakah seorang mahasiswa akan lulus <5 tahun atau ≥5 tahun. Hasil tersebut langsung digunakan dengan diperlihatkan secara langsung kepada pengguna setelah proses prediksi dilakukan.

4. Kesimpulan

Kesimpulan yang dapat diambil dari hasil penelitian ini adalah sebagai berikut:

1. Dari hasil penerapan algoritma C4.5 dalam prediksi kelulusan mahasiswa Prodi Informatika dapat disimpulkan bahwa atribut yang paling dominan dalam kelulusan mahasiswa adalah IPK, kedua adalah TOEFL, ketiga adalah asal daerah, dan yang terakhir adalah jenis kelamin.
2. Pemotongan pohon keputusan pada algoritma C4.5 dapat meningkatkan akurasi. Pohon tanpa pemotongan menghasilkan nilai rata-rata *precision* 63.93%, *recall* 60.73%, dan akurasi 60.52%. Sedangkan pohon keputusan yang dipotong dengan menggunakan metode *error-based pruning* dengan menggunakan nilai *confidence* 0,4 menghasilkan *precision* 70.70%, *recall* 50.65%, dan akurasi 61.57%. Pohon yang dipotong dengan menggunakan nilai *confidence* 0,25 menghasilkan *precision* 73.77%, *recall* 48.84%, dan akurasi 62.44%. Dari hasil tersebut dapat disimpulkan bahwa penggunaan nilai *confidence* 0,25 meningkatkan akurasi lebih baik daripada nilai *confidence* 0,4.

5. Daftar Pustaka

[1] S. D. Jadhav and H. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *International Journal of Science and Research (IJSR),* vol. 5, no. 1, pp. 1842-1845, January 2016.

[2] M. Mohankumar, S. Amuthakkani and G. Jeyamala, "Comparative Analysis of Decision Tree Algorithms for The Prediction of Eligibility of A Man for Availing Bank Loan," *International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST),* vol. 2, no. 15, pp. 360-366, 2016.

[3] D. H. Kamagi, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *ULTIMATICS,* vol. VI, no. 1, pp. 15-20, Juni 2014.

[4] V. Anggresta, "Analisis Faktor-faktor yang Mempengaruhi Belajar Mahasiswa Fakultas Ekonomi Universitas Negeri Padang," *Journal of Economic and Economic Education,* vol. 4, pp. 19-29, 2015.

[5] M. Devinta, "Fenomena Culture Shock (Gegar Budaya) pada Mahasiswa Perantauan di Yogyakarta," *Jurnal Pendidikan Sosiologi,* 2015.

[6] Y. Kusnia, "Pengaruh Karakteristik Gender dan Motivasi Belajar terhadap Prestasi Belajar Matematika Siswa Kelas X IPA 1 di MAN 2 Semarang," 2017.

[7] J. E. Carson, P. L. Carrell, S. Silberstein, B. Kroll and P. A. Kuehn, "Reading-Writing Relationships in First and Second Language," *TESOL Quarterly,* vol. 24, pp. 245-266, 1990.

[8] A. Saifudin and R. S. Wahono, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *Journal of Software Engineering,* vol. 1, pp. 76-85, 2015.

[9] K. Grabczewski, *Meta-Learning in Decision Tree Induction*, Springer, 2014.

[10] J. R. Quinlan, C4.5: *Programs for Machine Learning*, San Mateo: Morgan Kauffman, 1993.

[11] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Waltham: Elsevier Inc., 2014.