MODEL ANALISIS CLASSIFICATION UNTUK DATA MAHASISWA DAN DOSEN DI PERGURUAN TINGGI

by Mewati Ayub, Tanti Kristanti Maresha Caroline, Tjio Marvin Christian

Submission date: 12-Aug-2021 12:38PM (UTC+0700)
Submission ID: 1630517698
File name: MODEL_ANALISIS_CLASSIFICATION_UNTUK_DATA_MAHASISWA_DAN_DOSEN.pdf (901.72K)
Word count: 4886
Character count: 27738

MODEL ANALISIS *CLASS*IFICATION UNTUK DATA MAHASISWA DAN DOSEN DI PERGURUAN TINGGI

Mewati Ayub[1)], Tanti Kristanti[2)], Maresha Caroline[3)], Tjio Marvin Christian[4)]
[1,2,3,4] Fakultas Teknologi Informasi, Universitas Kristen Maranatha
email: mewati.ayub@itmaranatha.org, tanti.kristanti@itmaranatha.org, maresha.caroline@itmaranatha.org, t.marvin.christian@gmail.com

Abstract

The existence of historical data in universities, for instance, the data of lecturer and student academic process, is very valuable because of the fact that historical data can be analyzed to extract the implicit knowledge with the use of various analysis methods, which in turn can be used as a basis for improving education system. This research is aimed to analyze the historical data of students and lecturers using predictive classification methods as a continuation from previous research that had been managed to produce a data warehouse schema for both types of data. In addition, the research is focused on three datasets, which are datasets of research and community services and also the dataset of graduates that were extracted using the classification methods of decision tree, especially J48 with some confidence factor parameter settings as well as several minimum numbers of instances for the leaves that will produce optimal analysis model. Research methodology began with the formation of datasets derived from data star schema of the previous research results, followed by the determination of attributes of the datasets that would be used as training data and test data. Furthermore, those datasets were analyzed using some classification models before they finally were evaluated. The results of some test cases indicated that the decreasing in the value of the confidence factor and also the increasing of minimum number of instances on leaves in J48 classification, both were affect the resulting tree pruning. For research and community services datasets, the convergence on the number of leaves in the tree would be more quickly achieved along with the increase of minimum value of instances when compared to the case of the decrease of confidence factor value. Meanwhile, for the graduate dataset, in cases of different classes, the convergence of the leaves number was influenced by the distribution of the data in the class attribute, either by adding the minimum value of the instance on the leaves as well as by lowering the confidence factor value.

Keywords/Kata kunci: data mining, classification, decision tree, J48, universities historical data.

1. Pendahuluan

Penelitian Ranjan (Ranjan dan Khalil, 2008) menyebutkan bahwa pengetahuan yang diperoleh dari *data mining* dapat dimanfaatkan untuk meningkatkan kualitas pendidikan. Berdasarkan data histori yang dimiliki institusi, dapat dilakukan pengumpulan data yang dilanjutkan dengan analisis menggunakan *classification* sebagai metode prediktif untuk mengevaluasi data histori. Dari hasil analisis tersebut, dapat diperoleh pengetahuan yang sangat bermanfaat untuk merencanakan perbaikan dalam sistem pendidikan.

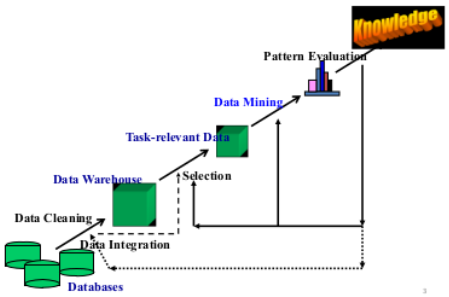
Berbagai penelitian *data mining* telah dilakukan untuk melakukan analisis terhadap data histori yang dimiliki institusi. Dalam penelitian (Bhardwaj dan Pal, 2014) digunakan klasifikasi dengan Naïve Bayes untuk memprediksi kinerja akademik mahasiswa. Sedangkan penelitian (Radaideh dan Al Nagi, 2012) menggunakan klasifikasi ID3, C4.5 dan Naïve Bayes untuk memprediksi kinerja karyawan. (Gibert, Maare, dan Codina 2014) juga memanfaatkan klasifikasi untuk analisis data yang berhubungan dengan masalah lingkungan hidup.

Dari penelitian tahun pertama (Ayub, Kristanti, dan Caroline, 2013), telah dihasilkan skema *data warehouse* untuk data mahasiswa dan data dosen yang dapat diterapkan di perguruan tinggi. Berbasis skema *data warehouse* yang dihasilkan, pada penelitian tahun kedua ini, akan dilakukan analisis data dengan klasifikasi untuk menghasilkan model analisis terhadap data dosen dan data mahasiswa. Studi kasus analisis *data mining* akan dilakukan terhadap *dataset* penelitian dosen, *dataset* pengabdian dosen, dan *dataset* lulusan. Metode klasifikasi yang digunakan adalah *decision tree* dengan J48 dari Weka yang dilengkapi dengan pengaturan parameter *confidence factor* dan jumlah minimal *instance* pada daun untuk dapat menghasilkan model analisis yang optimal. Metode J48 merupakan implementasi metode klasifikasi C4.5 yang paling berpengaruh dalam penelitian *data mining* (Wu, et al., 2008).

2. Landasan Teori

2.1 Data Mining

Data mining dikenal juga dengan istilah *Knowledge Discovery in Databases* (KDD). KDD merupakan prosedur yang bersifat interaktif dan iteratif yang berusaha mengekstraksi pengetahuan yang belum diketahui (implisit) menjadi pengetahuan yang bermanfaat dari sekumpulan data. Tahapan yang terjadi dalam KDD dapat dilihat pada Gambar 1.



Gambar 2. *Knowledge Discovery in Databases* (Han, Kamber, dan Pei, 2012)

Berdasarkan tujuan pemanfaatannya, *data mining* dapat dibedakan atas:

1. Metode prediktif

Dalam metode ini, *data mining* menggunakan beberapa variabel/parameter untuk memprediksi nilai variabel lain yang belum diketahui. Teknik *data mining* yang termasuk metode prediksi antara lain *classification* dan regresi.

2. Metode deskriptif

Dalam metode ini, *data mining* digunakan untuk menggambarkan karakteristik dari sekumpulan data dalam *dataset*, juga untuk menggambarkan pola pengetahuan yang tersembunyi di dalam kumpulan data tersebut. Teknik *data mining* yang termasuk metode deskripsi adalah *clustering* dan aturan asosiasi.

2.2 Classification

Pada teknik *classification*, diberikan sekumpulan *dataset*, setiap *dataset* akan terdiri dari beberapa atribut, salah satu dari atribut tersebut akan menentukan kelas (*class*). Klasifikasi akan mencari suatu model untuk atribut kelas sebagai fungsi dari nilai atribut lainnya. Berdasarkan model tersebut, suatu data baru harus dapat ditempatkan ke dalam suatu kelas setepat mungkin.

Dalam membangun model tersebut, akan digunakan kumpulan *dataset* yang disebut sebagai data *training.* Berdasarkan data *training,* fungsi untuk menentukan atribut kelas akan dirumuskan. Untuk menjamin keakuratan model tersebut, akan digunakan sekumpulan *dataset* yang disebut data tes. Model akan diuji dengan data tes, kemudian ditinjau persentase keberhasilannya di dalam mengklasifikasikan data tes.

Classification yang akan digunakan dalam penelitian ini adalah dengan pohon keputusan. Pohon keputusan (*decision tree*) digambarkan dengan suatu struktur pohon, dimana setiap *node* menyatakan pemeriksaan terhadap suatu atribut, setiap cabang menyatakan hasil (*outcome*) dari pemeriksaan tersebut, dan *node* daun menyatakan kelas. Pada saat mengklasifikasi sampel yang belum diketahui, nilai atribut sampel diperiksa terhadap pohon keputusan. Suatu jalur (*path*) ditelusuri dari akar sampai ke suatu *node* daun, sehingga diketahui kelas dari sampel tersebut.

Pada saat dilakukan *mining* data, terdapat dua kelompok data, yaitu *data training* dan *data test*. *Data training* digunakan untuk membentuk model analisis melalui suatu teknik *data mining*. Sedangkan *data test* digunakan untuk mengukur sejauh mana akurasi dari model analisis tersebut. Dengan demikian, hasil analisis yang diperoleh dari suatu teknik *data mining* perlu dievaluasi akurasinya dengan mengukur tingkat kesalahan (*error rate*) dari model yang dihasilkan (Witten, Frank, dan Hall, 2011) (Han, Kamber, dan Pei, 2012).

Untuk memprediksi performansi *classifier* pada data baru, perlu dilakukan dahulu penilaian *error rate* pada sebuah *dataset* yang bersifat independen dan tidak ada kaitannya dengan formasi *classifier*. *Dataset* yang independen tersebut disebut juga *test set*. Baik *training data* maupun *test data* merupakan contoh-contoh dari permasalahan yang akan dianggap penting dan representatif (Witten, Frank, dan Hall, 2011).

Umumnya, semakin besar sampel *training*, akan semakin baik *classifier*-nya, meskipun hasilnya berkurang ketika volume data *training* berlebih. Semakin besar sampel *test*, maka semakin akurat estimasi *error*-nya. Masalah akan muncul ketika data yang tersedia tidaklah besar. Pada situasi seperti ini, data *training* harus diklasifikasi secara manual dan begitu pula data *test* harus mengandung estimasi *error*. Hal tersebut akan membatasi jumlah data yang akan digunakan untuk *training, validation,* dan *testing*.

Jika semua contoh dengan *class* tertentu dihilangkan dari data *training*, maka data *classifier* akan dipakai untuk melakukan pengujian pada *class* tersebut dan hal tersebut akan diperburuk dengan fakta bahwa *class* perlu ditempatkan pada data *test* karena tidak ada satu pun *instance*-nya yang diperlukan untuk data *training*. Namun, harus dipastikan

bahwa *random sampling* dilakukan sedemikian sehingga setiap *class* disiapkan untuk *training set* dan *test set* (Witten, Frank, dan Hall, 2011).

Cara umum untuk mengurangi setiap bias yang disebabkan karena sampel tertentu adalah dengan cara mengulangi seluruh proses, *training and testing* berulang kali dengan *random sample* yang berbeda. Dalam setiap iterasi, dapat digunakan misalnya dua per tiga data untuk *training*, dimungkinkan dengan *stratification,* dan sisanya dapat digunakan untuk *testing*. *Error rate* pada setiap iterasi yang berbeda akan dirata-rata untuk menghasilkan keseluruhan *error rate*. Prosedur ini disebut *repeated holdout method* untuk melakukan estimasi *error rate*.

Namun, dapat juga digunakan teknik statistik yaitu *cross-validation*. Dalam *cross-validation,* dapat dipilih sebuah *fixed number* untuk *fold* atau partisi dari data tersebut. Sebagai contoh jika memakai tiga partisi, maka data akan dibagi tiga dengan pembagian yang hampir sama, sepertiga data akan digunakan untuk *testing* dan sisanya digunakan untuk *training* dan prosedurnya akan diulangi sebanyak tiga kali sehingga pada akhir proses, setiap *instance* akan digunakan sekali untuk pengujian. Prosedur ini disebut *threefold cross-validation* (Witten, Frank, dan Hall, 2011).

Standar yang digunakan untuk memprediksi *error rate* pada klasifikasi dengan *single, fixed sample data* adalah *stratified tenfold cross-validation* dimana data dibagi secara acak menjadi sepuluh bagian dimana *class* direpresentasikan dalam *full dataset* dengan proporsi yang hampir sama. Setiap bagian akan diuji dan *error rate*-nya akan dikalkulasi. Dengan demikian, klasifikasi dieksekusi sebanyak sepuluh kali pada *training set* yang berbeda (setiap set memiliki kesamaan). Terakhir, kesepuluh *error estimate* yang didapatkan akan dirata-rata untuk menghasilkan keseluruhan *error estimate*. Angka sepuluh dipilih karena berdasarkan sejumlah bukti teoritis yang melatarbelakanginya, angka ini dianggap nilai *fold* yang tepat untuk mendapatkan *error estimate* yang terbaik. Meskipun argumen ini masih menuai sejumlah perdebatan, namun *tenfold cross-validation* telah menjadi metoda yang standar dilakukan. Sejumlah pengujian juga telah menunjukkan bahwa penggunaan stratifikasi akan meningkatkan hasil sedikit demi sedikit. Oleh karenanya teknik evaluasi standar dalam situasi dimana data sangat terbatas akan memerlukan *stratified tenfod cross-validation*. Pembagian data untuk stratifikasi dan juga 10 *fold* tidak perlu sama, cukup dengan membaginya menjadi sepuluh set dengan ukuran yang hampir sama dengan representasi keragaman nilai-nilai *class* yang hampir sama, jadi dimungkinkan memiliki perbandingan 10:5 *fold* atau 20-*fold cross-validation.* Stratifikasi mengurangi variasi, namun tidak mengurangi keseluruhan data (Witten, Frank, dan Hall, 2011).

Selain *training error*, model hasil klasifikasi juga ditentukan oleh *generalization error*. *Generalization error* adalah kesalahan yang ditimbulkan oleh data baru yang tidak muncul dalam proses klasifikasi. Model klasifikasi yang baik tidak hanya dapat memetakan data *training* dengan tepat, tetapi juga harus dapat mengklasifikasi data baru yang tidak muncul dalam data *training*. Apabila suatu model klasifikasi memetakan data *training* dengan sangat tepat, sehingga tidak dapat mengklasifikasi data baru dengan benar, maka model tersebut bersifat *overfitting* (Tan, Steinbach, dan Kumar, 2006).

Overfitting dapat diatasi dengan melakukan pemangkasan (*pruning*) terhadap *tree* hasil klasifikasi. Terdapat dua macam pemangkasan, yaitu *pre-pruning* dan *post-pruning*. Model *tree* hasil klasifikasi yang telah mengalami pemangkasan, bersifat lebih umum (*general*) dalam mengklasifikasi data baru, sehingga kesalahan *generalisasi* dapat diperkecil (Tan, Steinbach, dan Kumar, 2006) (Witten, Frank, dan Hall, 2011).

3. Metode Penelitian

Tahapan yang dilakukan dalam penelitian adalah sebagai berikut:

a. Membentuk *dataset* yang berasal dari skema data yang berbentuk *star* hasil penelitian tahun pertama.
b. Menentukan atribut *dataset* yang akan digunakan sebagai data *training* dan data *test* dalam analisis dengan *classification.*
c. Melakukan analisis dengan model *classification* untuk *dataset* dengan parameter nilai *confidence factor* yang diubah-ubah untuk jumlah *minimum instance* tertentu pada daun menggunakan 10 *fold cross validation*.
d. Melakukan analisis dengan model *classification* untuk *dataset* dengan parameter nilai jumlah minimum *instance* pada daun yang diubah-ubah dengan nilai *confidence factor* tertentu menggunakan 10 *fold cross validation*.
e. Melakukan evaluasi terhadap model hasil analisis tersebut di atas.

Metodologi di atas digunakan untuk menganalisis *dataset* penelitian dosen, *dataset* pengabdian dosen, dan *dataset* lulusan.

4. Hasil Dan Pembahasan

Penelitian yang telah dilakukan adalah klasifikasi terhadap *dataset* penelitian dosen, *dataset* pengabdian dosen, dan *dataset* lulusan. Klasifikasi yang dilakukan menggunakan metode J48 yang merupakan implementasi algoritma klasifikasi C4.5 revisi ke-8 yang disediakan oleh Weka (Witten, Frank, dan Hall, 2011). Algoritma ini akan menggunakan *Gain Ratio* sebagai nilai *heuristic* untuk menentukan atribut penentu dalam *decision tree* yang dibentuk.

4.1 *Dataset* Penelitian Dosen

Dataset penelitian dosen terdiri dari 309 *instances*. Deskripsi dari *dataset* penelitian dosen dapat dilihat pada Tabel 1.

Tabel 1 Deskripsi *Dataset* Penelitian Dosen

No.	Nama Atribut	Keterangan
1.	IdFakultas	1 = 50 data 2 = 52 data 3 = 34 data 4 = 19 data 5 = 83 data 6 = 37 data 7 = 28 data 8 = 6 data
2.	IdJenjangPendidikan	3 = 34 data 4 = 236 data 5 = 39 data
3.	IdJabatanAkademik	1 = 116 data 2 = 98 data 3 = 40 data 4 = 5 data 5 = 50 data
4.	IdKelompokUsia	1 = 159 data 2 = 90 data 3 = 60 data
5.	IdKelompokMasa Kerja	1 = 152 data 2 = 125 data 3 = 32 data
6.	Jumlah Riset (*atribute class*)	Low = 167 data Med = 87 data High = 55 data

Gain Ratio yang diperoleh berdasarkan data pada Tabel 1 dan sudah diurutkan dari nilai tertinggi sampai nilai terendah dapat dilihat pada Tabel 2.

Tabel 2 Nilai *Gain Ratio Dataset* Penelitian Dosen

No.	Nama Atribut	*Gain Ratio*
1.	IdKelompokMasaKerja	0.0491
2.	IdFakultas	0.0447
3.	IdJabatanAkademik	0.0389
4.	IdJenjangPendidikan	0.0249
5.	IdKelompokUsia	0.0178

Hasil klasifikasi dari *dataset* pada Tabel 1 dengan klasifikasi J48 dapat dilihat pada Tabel 3. Pada model awal yang dihasilkan dari percobaan dengan *confidence factor* = 0.25, empat atribut penentu dalam *tree* mengikuti urutan yang dihasilkan dari *gain ratio* pada Tabel 2. *Tree* awal yang dihasilkan bersifat *overfitting* karena menggambarkan data *training* yang digunakan. Sifat *overfitting* menyebabkan *tree* sangat spesifik dan model yang dihasilkan tidak bisa digunakan untuk generalisasi data (Tan, Steinbach, dan Kumar, 2006). Oleh karena itu dilakukan beberapa kali percobaan dengan nilai *confidence factor* yang diubah-ubah, mulai dengan nilai *confidence factor* = 0.25 sampai dengan 0.08. Nilai *confidence factor* yang mengecil menyebabkan dilakukan pemangkasan (*pruning*) terhadap *tree* yang dihasilkan. Pada akhir percobaan, diperoleh *tree* dengan jumlah daun 26. Pada *tree* akhir yang dihasilkan, hanya tiga atribut yang menjadi penentu dalam *tree*.

Tabel 3 Hasil Klasifikasi *Dataset* Penelitian Dosen dengan Variasi *Confidence Factor*

No.	*ConfidenceFactor*	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar(%)
1.	0.25	34	59.22
2.	0.20	34	58.90
3.	0.15	28	56.96
4.	0.10	28	57.28
5.	0.09	28	54.05
6.	0.08	26	54.05
7.	0.07	Tidak terbentuk *tree*	53.07

Selanjutnya dilakukan juga klasifikasi dengan nilai *Confidence Factor* tetap sebesar 0.25 dan ada variasi pada *minimum instance* pada daun, mulai dengan jumlah *instance* = 2 sampai dengan 30. Hasilnya dapat dilihat pada Tabel 4. Pada akhir percobaan diperoleh *tree* dengan jumlah daun = 3, adapun atribut penentu yang digunakan hanya satu.

Tabel 4 Hasil Klasifikasi *Dataset* Penelitian Dosen dengan *Confidence Factor* = 0.25 dan dengan Variasi *Minimum Instance* pada daun

No.	***Minimum Instance* pada Daun**	**Jumlah Daun dalam *Tree***	***Instance* yang terklasifikasi dengan benar (%)**
1.	2	34	59.22
2.	5	28	57.60
3.	10	18	57.28
4.	15	14	54.05
5.	20	3	54.37
6.	30	3	55.99
7.	40	Tidak terbentuk *tree*	

Karena IdFakultas merupakan atribut dengan 8 variasi nilai, maka untuk memperoleh hasil analisis yang lebih umum, percobaan selanjutnya dilakukan dengan menghilangkan atribut IdFakultas dari *dataset* pada Tabel 1. Hasil klasifikasi J48 untuk *dataset* penelitian dosen tanpa IdFakultas dapat dilihat pada Tabel 5. Dari percobaan tersebut diperoleh *tree* yang sama dengan *tree* yang diperoleh pada Tabel 4.

Tabel 5 Hasil Klasifikasi *Dataset* Penelitian Dosen tanpa IdFakultas

No.	***ConfidenceFactor***	**Jumlah Daun dalam *Tree***	***Instance* yang terklasifikasi dengan benar(%)**
1.	0.25	3	55.34
2.	0.20	3	55.34
3.	0.15	3	55.02
4.	0.10	Tidak terbentuk *tree*	

Dari model hasil penelitian pada Tabel 4 dan Tabel 5, *class: High* untuk atribut Jumlah Riset tidak muncul dalam *tree*, hal ini disebabkan distribusi data yang tidak seimbang untuk atribut Jumlah Riset, seperti tampak pada Tabel 1. Untuk mengatasi data *imbalance* sehingga *class: High* dapat muncul dalam *tree*, digunakan *SpreadSubSample* (Ganganwar, 2012) (Witten, Frank, dan Hall, 2011). Deskripsi datanya dapat dilihat pada Tabel 6.

Tabel 6 Deskripsi *Dataset* Penelitian Dosen tanpa IdFakultas dengan *SpreadSubSample*

No.	Nama Atribut	Keterangan
1.	IdJenjangPendidikan	3 : 12 data 4 : 133 data 5 : 20 data
2.	IdJabatanAkademik	1 : 63 data 2 : 59 data 3 : 22 data 4 : 3 data 5 : 18 data
3.	IdKelompokUsia	1 : 88 data 2 : 41 data 3 : 36 data
4.	IdKelompokMasa Kerja	1 : 71 data 2 : 71 data 3 : 23 data
5.	Jumlah Riset (atribut *class*)	Low : 55 data Med : 55 data High : 55 data

Dari percobaan yang dilakukan untuk Tabel 6, diperoleh *tree* dengan 3 daun, di mana semua *class* muncul dalam *tree*, dengan persentase *instance* yang terklasifikasi dengan benar sebesar 47.88%.

4.2 *Dataset* Pengabdian Dosen

Dataset pengabdian dosen terdiri dari 282 *instances*. Deskripsi dari *dataset* pengabdian dosen dapat dilihat pada Tabel 7.

Tabel 7 Deskripsi *Dataset* Pengabdian Dosen

No.	Nama Atribut	Keterangan
1.	IdFakultas	1 =61 data 2 =37 data 3 =24 data 4 =6 data 5 = 82 data 6 = 35 data 7 = 28 data 8 =9 data
2.	IdJenjangPendidikan	3 = 38 data 4 = 212 data 5 = 32 data
3.	IdJabatanAkademik	1 =99 data 2 = 93 data 3 =34 data 4 =3 data 5 = 53 data
4.	IdKelompokUsia	1 = 145 data 2 =69 data 3 = 68 data
5.	IdKelompokMasa Kerja	1 = 146 data 2 = 100 data 3 = 36 data
6.	Jumlah Riset (*atribute class*)	Low = 126 data Med =106 data High = 50 data

Gain Ratio yang terbentuk berdasarkan data pada Tabel 7 dan sudah diurutkan dari nilai tertinggi sampai nilai terendah dapat dilihat pada Tabel 8.

Tabel 8 Nilai *Gain Ratio Dataset* Pengabdian Dosen

No.	Nama Atribut	*Gain Ratio*
1.	IdJabatanAkademik	0.0392
2.	IdFakultas	0.0388
3.	IdJenjangPendidikan	0.0219
4.	IdKelompokMasaKerja	0.0154
5.	IdKelompokUsia	0.0146

Pada Tabel 9 berisi hasil klasifikasi J48 terhadap *dataset* dari Tabel 7 dengan variasi *Confidence Factor*, mulai dari 0.25 sampai dengan 0.09. Pada model awal yang dihasilkan dari percobaan dengan *confidence factor* = 0.25, empat atribut penentu dalam *tree* mengikuti urutan yang dihasilkan dari *gain ratio* pada Tabel 8. Pada akhir percobaan diperoleh model dengan jumlah daun 22 dengan tiga atribut penentu dalam *tree*.

Tabel 9 Hasil Klasifikasi *Dataset* Pengabdian Dosen dengan Variasi *Confidence Factor*

No.	*ConfidenceFactor*	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar(%)
1.	0.25	37	44.68
2.	0.20	32	45.39
3.	0.15	29	46.80
4.	0.10	22	44.68
5.	0.09	22	43.97
6.	0.08	Tidak terbentuk *tree*	

Sesuai dengan percobaan yang dilakukan pada *dataset* penelitian dosen, maka *dataset* pengabdian dosen juga diklasifikasikan dengan nilai *Confidence Factor* tetap sebesar 0.25 dan adanya variasi pada *minimum instance* pada daun, mulai dengan jumlah *instance* = 2 sampai dengan 90. Hasilnya dapat dilihat pada Tabel 10. Pada model akhir yang dihasilkan, terdapat satu atribut penentu dalam *tree*.

Tabel 10 Hasil Klasifikasi *Dataset* Pengabdian Dosen dengan *Confidence Factor* = 0.25 dan dengan Variasi *Minimum Instance* pada daun

No.	*Minimum Instance* pada Daun	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar (%)
1.	2	37	44.68
2.	5	26	46.80
3.	10	19	47.16
4.	15	12	44.68
5.	20	12	44.32
6.	30	5	44.68
7.	50	5	45.03
8.	70	5	44.33
9.	90	5	44.68

Selain itu dilakukan juga percobaan dengan menghilangkan IdFakultas dari *dataset* pengabdian dosen, hasilnya dapat dilihat pada Tabel 11. Dari percobaan tersebut diperoleh *tree* yang sama dengan *tree* yang diperoleh pada Tabel 10.

Tabel 11 Hasil Klasifikasi *Dataset* Pengabdian Dosen tanpa IdFakultas

No.	*Confidence Factor*	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar(%)
1.	0.25	24	41.84
2.	0.20	16	40.78
3.	0.15	5	42.20
4.	0.10	Tidak terbentuk tree	44.68

4.3 *Dataset* Lulusan

Dataset lulusan terdiri dari 2562 *instances*. Deskripsi dari data lulusan dapat dilihat pada Tabel 12.

Tabel 12 Deskripsi *Dataset* Lulusan

No.	Nama Atribut	Keterangan
1.	IdGelombang	1 = 1394 data 2 = 844 data 3 = 324 data
2.	IdKelompokNilaiUSM	1 = 487 data 2 = 679 data 3 = 461 data 4 = 827 data 5 = 108 data
3.	IdJurusanSMA	1 = 1690 data 2 = 872 data
4.	IdWilayah	1 = 441 data 2 = 2121 data
5.	SksTempuh	1 = 1950 data 2 = 26 data 3 = 586 data
6.	SksLulus	1 = 1340 data 2 = 977 data 3 = 245 data
7.	LamaStudi	1 = 1454 data 2= 1108 data
8.	IdKelompokIP (*atribute class*)	1 = 853 data 2 = 1271 data 3 = 438 data

Klasifikasi untuk *dataset* Lulusan, dilakukan dalam dua kelompok percobaan, yaitu kelompok percobaan dengan IdKelompokIP sebagai *class* dan kelompok percobaan dengan LamaStudi sebagai *class*.

Gain Ratio untuk kelompok percobaan dengan IdKelompokIP sebagai *class* yang terbentuk berdasarkan data pada Tabel 12 dan sudah diurutkan dari nilai tertinggi sampai nilai terendah dapat dilihat pada Tabel 13.

Tabel 13 Nilai *Gain Ratio Dataset* Lulusan (Class IdKelompokIP)

No.	Nama Atribut	*Gain Ratio*
1.	LamaStudi	0.25817
2.	IdKelompokNilaiUSM	0.15638
3.	SKSLulus	0.1515
4.	IdJurusanSMA	0.09703
5.	IdGelombang	0.0363
6.	IdWilayah	0.00576
7.	SKSTempuh	0.00522

Hasil klasifikasi J48 untuk *dataset* Lulusan dengan IdKelompokIP sebagai *class* dapat dilihat pada Tabel 14.

Tabel 14 Hasil Klasifikasi *Dataset* Lulusan dengan Variasi *Confidence Factor*

No.	*ConfidenceFactor*	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar(%)
1.	0.25	46	69.63
2.	0.20	33	69.28
3.	0.15	26	69.28
4.	0.10	26	69.47
5.	0.05	22	69.20
6.	0.01	17	67.29

Sesuai dengan percobaan yang dilakukan pada *dataset* dosen, maka *dataset* lulusan juga diklasifikasikan dengan nilai *Confidence Factor* sebesar 0.25 dan adanya variasi pada *minimum instance* pada daun. Hasilnya dapat dilihat pada Tabel 15.

Tabel 15 Hasil Klasifikasi *Dataset* Lulusan dengan *Confidence Factor* = 0.25 dan dengan Variasi *Minimum Instance* pada daun

No.	*Minimum Instance* pada Daun	Jumlah Daun dalam *Tree*	*Instance* yang terklasifikasi dengan benar (%)
1.	2	46	69.63
2.	5	34	69.24
3.	10	30	69.16
4.	15	30	69.32
5.	20	29	69.24
6.	30	21	69.24
7.	40	18	69.51
8.	50	18	69.63
9.	60	18	69.63
10.	70	18	69.63
11.	80	18	68.85
12.	90	10	68.11
13.	100	10	68.11

Gain Ratio untuk *dataset* pada Tabel 12 dengan LamaStudi sebagai *class* ditunjukkan pada Tabel 16 yang diurutkan berdasarkan nilai *gain ratio* dari nilai tertinggi sampai dengan nilai terendah.

Tabel 16 Nilai *Gain Ratio Dataset* Lulusan dengan LamaStudi sebagai *Class*

No.	Nama Atribut	*Gain Ratio*
1.	IdKelompokIP	0.173823
2.	IdKelompokNilaiUSM	0.080082
3.	SKSLulus	0.053162
4.	IdJurusanSMA	0.053105
5.	SKSTempuh	0.033141
6.	IdGelombang	0.019504
7.	IdWilayah	0.000963

Hasil klasifikasi J48 terhadap *dataset* Lulusan dengan *class* LamaStudi dapat dilihat pada Tabel 17 dengan variasi nilai *confidence factor* mulai dari 0.25 sampai dengan 0.01.

Tabel 17 Hasil Klasifikasi *Dataset* Lulusan (*Class* LamaStudi) dengan Variasi *Confidence Factor*

No.	*Confidence Factor*	Jumlah Daun dalam *tree*	*Instance* yang terklasifikasi dengan benar (%)
1.	0.25	9	79.43
2.	0.20	5	79.62
3.	0.15	5	79.62
4.	0.10	5	79.62
5.	0.05	5	79.66
6.	0.01	5	79.66

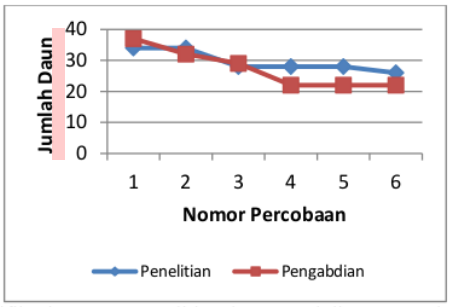
Hasil klasifikasi J48 terhadap *dataset* Lulusan dengan *class* Lama Studi dapat dilihat pada Tabel 18 dengan *confidence factor* = 0.25 dan jumlah *minimum instance* pada daun bervariasi dari 2 daun sampai dengan 100 daun.

Tabel 18 Hasil Klasifikasi *Dataset* Lulusan (*Class* LamaStudi) dengan *Confidence Factor* = 0.25 dan *Minimum Instance* pada Daun Bervariasi

No.	*Minimum Instance* pada Daun	Jumlah Daun Dalam *Tree*	*Instance* yang terklasifikasi dengan benar (%)
1.	2	9	79.43
2.	5	9	79.58
3.	10	5	79.58
4.	15	5	79.58
5.	20	5	79.58
6.	30	5	79.66
7.	40	5	79.66

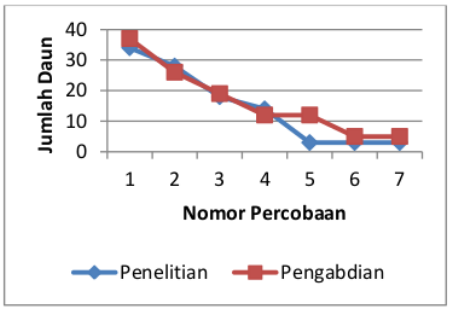
4.4 Hasil Analisis

Pada Gambar 2 ditampilkan grafik yang menunjukkan hasil percobaan pada Tabel 3 dan Tabel 9, yaitu klasifikasi *dataset* Penelitian Dosen dan *dataset* Pengabdian Dosen dengan variasi nilai *confidence factor* pada nilai *minimum instance* pada daun = 2.



Gambar 2 Hasil Percobaan Klasifikasi *Dataset* Penelitian dan Pengabdian Dosen dengan Variasi *Confidence Factor*

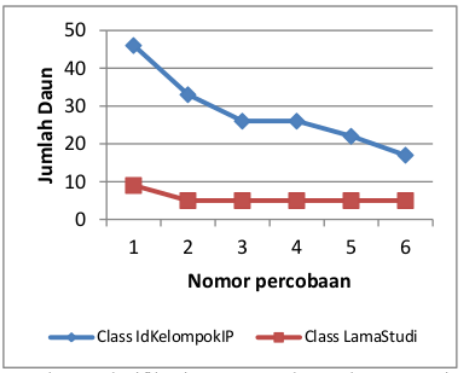
Pada Gambar 3 ditampilkan grafik yang menunjukkan hasil percobaan pada Tabel 4 dan Tabel 10, yaitu klasifikasi *dataset* Penelitian Dosen dan *dataset* Pengabdian Dosen dengan variasi nilai *minimum instance* pada daun untuk *confidence factor* = 0.25.



Gambar 3 Hasil Percobaan Klasifikasi *Dataset* Penelitian dan Pengabdian Dosen dengan Variasi *Minimum Instance* pada Daun

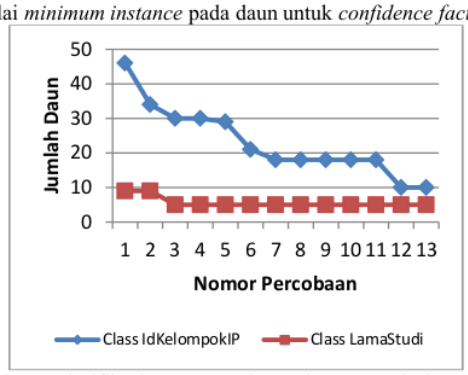
Dari Gambar 2, terlihat bahwa penurunan jumlah daun berhenti pada percobaan ke-6, yaitu pada nilai *confidence factor* = 0.08. Sedangkan pada Gambar 3, terlihat bahwa penurunan jumlah daun berhenti pada percobaan ke-7, yaitu pada nilai *minimum instance* pada daun = 40. Dari perbandingan grafik pada Gambar 2 dan Gambar 3, dapat disimpulkan bahwa penambahan nilai *minimum instance* pada daun menyebabkan penurunan jumlah daun dalam *tree* yang dihasilkan lebih cepat dibandingkan penurunan nilai *confidence factor*. Dengan demikian *tree* yang dihasilkan dari penambahan nilai *minimum instance* lebih ramping dan bersifat lebih umum (*general*) dibandingkan *tree* yang dihasilkan dari penurunan *confidence factor*.

Pada Gambar 4 ditampilkan grafik yang menunjukkan hasil percobaan pada Tabel 14 dan Tabel 17, yaitu klasifikasi *dataset* Lulusan dengan variasi nilai *confidence factor* pada nilai *minimum instance* = 2.



Gambar 4 Hasil Percobaan Klasifikasi *Dataset* Lulusan dengan Variasi *Confidence Factor*

Pada Gambar 5 ditampilkan grafik yang menunjukkan hasil percobaan pada Tabel 15 dan Tabel 18, yaitu klasifikasi *dataset* Lulusan dengan variasi nilai *minimum instance* pada daun untuk *confidence factor* = 0.25.



Gambar 5 Hasil Percobaan Klasifikasi *Dataset* Lulusan dengan Variasi *Minimum Instance* pada Daun

Dari Gambar 4, terlihat bahwa penurunan jumlah daun untuk *Class* IdKelompokIP berhenti pada percobaan ke-6, yaitu pada nilai *confidence factor* = 0.01. Sedangkan untuk *Class* LamaStudi, jumlah daun pada *tree* yang dihasilkan mulai konvergen pada percobaan ke-3, yaitu pada nilai *confidence factor* = 0.15.

Pada Gambar 5, terlihat bahwa penurunan jumlah daun untuk *Class* IdKelompokIP konvergen pada percobaan ke-13, yaitu untuk nilai *minimum instance* pada daun = 100. Sedangkan untuk *Class* LamaStudi, jumlah daun pada *tree* yang dihasilkan mulai konvergen pada percobaan ke-4, yaitu untuk nilai *minimum instance* pada daun = 15.

Dari perbandingan grafik pada Gambar 4 dan Gambar 5, dapat disimpulkan bahwa konvergensi jumlah daun dalam *tree* yang dihasilkan terjadi lebih cepat untuk klasifikasi dengan *Class* LamaStudi dibandingkan klasifikasi dengan *Class* IdKelompokIP. Distribusi data dalam *Class* LamaStudi pada Tabel 12 lebih merata dibandingkan dengan distribusi data dalam *Class* IdKelompokIP.

5. Kesimpulan

Kesimpulan yang dapat ditarik dari hasil penelitian adalah:

1. Penurunan nilai *confidence factor* dalam klasifikasi dengan J48 berpengaruh dalam pemangkasan (*pruning*) *tree* yang dihasilkan.
2. Penambahan nilai *minimum instance* pada daun dalam klasifikasi dengan J48 berpengaruh dalam pemangkasan (*pruning*) *tree* yang dihasilkan.
3. Dari percobaan terhadap *dataset* penelitian dosen dan *dataset* pengabdian dosen, konvergensi jumlah daun dalam *tree* yang dihasilkan dengan cara penambahan nilai *minimum instance* pada daun lebih cepat dicapai dibandingkan dengan cara penurunan nilai *confidence factor*.
4. Dari percobaan terhadap *dataset* Lulusan dengan atribut *class* yang berbeda, distribusi data dalam atribut *class* berpengaruh terhadap konvergensi jumlah daun dalam *tree* yang dihasilkan, baik dengan cara penambahan nilai *minimum instance* pada daun, maupun dengan cara penurunan nilai *confidence factor*.

Ucapan Terima Kasih

Terima kasih atas Hibah penelitian yang diberikan oleh DIPA Kopertis Wilayah IV, Kementrian Pendidikan dan Kebudayaan melalui LPPM Universitas Kristen Maranatha untuk tahun anggaran 2014.

6. Daftar Pustaka

[1] Ayub, M, Kristanti, T., dan Caroline, M. (2013). Data warehouse sebagai basis analisis data akademik perguruan tinggi, *Prosiding Seminar Nasional Teknologi Informasi*, Fakultas Teknologi Informasi Universitas Tarumanegara, 18 – 25.

[2] Bhardwaj, B.K. dan Pal, S. Data mining : A prediction for performance improvement using classification. http://arxiv.org/ftp/arxiv/papers/1201/1201.3418.pdf, diakses terakhir tanggal 6 Maret 2014.

[3] Han, J., dan Kamber, M., Pei, J. (2012). *Data mining concepts and techniques.* Edisi ke-3. Waltham: Morgan Kaufmann Publisher.

[4] Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4), 42 - 47.

[5] Gibert, K., Marre, M.S., dan Codina V. Choosing the right data mining technique : Classification of methods and intelligent recommendation, http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings, diakses terakhir tanggal 20 Februari 2014.

[6] Radaideh, Q.A. dan Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employees performance. *International Journal of Advanced Computer Science and Applications,* 3(2), 144-151.

[7] Ranjan, J. dan Khalil, S.(2008). Conceptual framework of data mining process in management education in India : An institutional perspective. *Information Technology Journal,* 7(1), 16-23.

[8] Tan P., Steinbach, M. dan Kumar, V. (2006). *Introduction to data mining*. Boston : Pearson International Ed.

[9] Witten, I.H., Frank, E., dan Hall, M.A. (2011). *Data mining practical machine learning tools and techniques.* Edisi ke-3. Burlington: Morgan Kaufmann Publisher.

[10] Wu, X., & et al. (2008). Top 10 algorithms in data mining. *Knowledge Information System Volume 14*, 1-37.

MODEL ANALISIS CLASSIFICATION UNTUK DATA MAHASISWA DAN DOSEN DI PERGURUAN TINGGI

ORIGINALITY REPORT

6%	6%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.its.ac.id Internet Source	1%
2	repositori.uin-alauddin.ac.id Internet Source	1%
3	Submitted to AUT University Student Paper	1%
4	eprints.unsri.ac.id Internet Source	1%
5	www.wwjournal.ir Internet Source	1%
6	dokumen.tips Internet Source	1%
7	pt.scribd.com Internet Source	1%
8	www.scielo.org.mx Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		