Volume 10 No. 1
Januari 2018
ISSN : 2085 – 1669
e-ISSN : 2460 – 0288
Website : jurnal.umj.ac.id/index.php/jurtek
Email : jurnalteknologi@umj.ac.id

Jurnal Teknologi
UNIVERSITAS MUHAMMADIYAH JAKARTA

METODE DATA MINING UNTUK SELEKSI CALON MAHASISWA PADA PENERIMAAN MAHASISWA BARU DI UNIVERSITAS PAMULANG

Aries Saifudin[1]
[1]Jurusan Teknik Informatika, Fakultas Teknik, Universitas Pamulang, Tangerang
Jl. Puspiptek No.23, Buaran, Serpong, Kota Tangerang Selatan, Banten 15310
Email: aries.saifudin@gmail.com

Diterima: 20 April 2017 Direvisi: 9 Juni 2017 Disetujui: 21 Juli 2017

ABSTRAK

Universitas Pamulang berusaha memberikan pendidikan tinggi dengan biaya yang terjangkau oleh kalangan bawah. Tetapi mahasiswanya banyak yang keluar di tiap semester, sehingga menyebabkan rasio jumlah mahasiswa baru dengan jumlah yang lulus tidak seimbang. Selain itu banyak mahasiswa yang tidak lulus tepat waktu, hal ini mengakibatkan rasio dosen dan mahasiswa tidak seimbang. Kedua hal ini akan mengurangi penilaian pada saat akreditasi. Penyebab keluarnya mahasiswa tanpa menyelesaikan pendidikannya, atau tidak dapat menyelesaikan pendidikannya tepat waktu belum dapat dideteksi dengan sistem seleksi saat ini. Pada penelitian ini diusulkan penggunaan teknik data mining untuk memprediksi ketepatan waktu lulus calon mahasiswa. Teknik data mining dan machine learning dapat digunakan untuk memprediksi berdasarkan data-data masa lalu. Metode *data mining* yang digunakan untuk memprediksi adalah klasifikasi, yaitu *Naïve Bayes* (NB), *k-Nearest Neighbor* (k-NN), *Random Forest* (RF), *Decision Stump* (DS), *Decision Tree* (DT), *Rule Induction (RI), Linear Regression* (LR), *Linear Discriminant Analysis* (LDA), *Neural Network* (NN), dan *Support Vector Machine* (SVM). Berdasarkan hasil implementasi dan pengukuran algoritma/model yang diusulkan diperoleh algoritma/model terbaik, yaitu *Support Vector Machine* (SVM) dengan akurasi 65.00%. Tetapi akurasi ini masih jauh dari nilai *excellent* (sangat baik).

Kata kunci: *Data Mining, Klasifikasi, Penerimaan Mahasiswa*

ABSTRACT

Pamulang University strives to provide higher education at an affordable cost by the lower classes. But many students are dropout in each semester, thus causing the ratio of the number of new students by the number of graduating unbalanced. In addition, many students who do not graduate on time, this resulted in the ratio of lecturers and students are not balanced. Both of these will reduce the assessment at the time of accreditation. The cause of the release of the student without completing their education, or unable to complete their education on time can not be detected with the current selection system. In this study, proposed to use data mining techniques to predict the timeliness graduate students. Data mining techniques and machine learning can be used to predict based on past data. The data mining method used to predict is the classification, namely Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Random Forest (RF), Decision Stump (DS), Decision Tree (DT), Rule Induction (RI) , Linear Regression (LR), Linear Discriminant Analysis (LDA), Neural Network (NN), and Support Vector Machine (SVM). Based on the results of the implementation and measurement algorithms/models proposed obtained the best algorithm/model, namely Support Vector Machine (SVM) with an accuracy of 65.00%. But this accuracy is still far from excellent value.

Keywords*: Data Mining, Classification, Student Enrollment*

DOI: https://dx.doi.org/10.24853/jurtek.10.1.25-36

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

PENDAHULUAN

Pendidikan tinggi di Indonesia telah mengalami peningkatan dari waktu ke waktu, saat ini sudah mencapai 3098 perguruan tinggi baik negeri maupun swasta (Wahyudin, 2015). Peningkatan juga terjadi pada jumlah pendaftar di perguruan tinggi (Yasmiati, Wahyudi, & Susilo, 2017). Peningkatan ini harus diimbangi dengan strategi seleksi penerimaan mahasiswa yang baik agar mendapatkan calon mahasiswa yang berkualitas, karena daya tampung dan tingkat kelulusan merupakan bagian penting dalam pengambilan keputusan (Bisri & Wahono, 2015). Mahasiswa yang berkualitas bukan hanya yang memiliki kemampuan intelektual tinggi, tetapi juga harus dapat menyelesaikan pendidikannya dengan baik.

Universitas Pamulang memiliki jumlah mahasiswa yang sangat banyak, tiap semester menerima mahasiswa baru dengan jumlah yang sangat banyak. Hal ini terjadi karena Universitas Pamulang berusaha memberikan pendidikan tinggi dengan biaya yang terjangkau oleh kalangan bawah. Tetapi mahasiswanya banyak yang keluar di tiap semester, sehingga menyebabkan rasio jumlah mahasiswa baru dengan jumlah yang lulus tidak seimbang. Selain itu banyak mahasiswa yang tidak lulus tepat waktu, hal ini mengakibatkan rasio dosen dan mahasiswa tidak seimbang. Kedua hal ini akan mengurangi penilaian pada saat akreditasi.

Mahasiswa memiliki tingkat motivasi yang berbeda, sikap yang berbeda dalam belajar, dan respon yang berbeda terhadap instruksi praktis khusus (Manhães, Cruz, & Zimbrão, 2014, p. 2). Banyaknya mahasiswa yang keluar sebelum menyelesaikan pendidikannya, dan yang tidak dapat menyelesaikan pendidikan tepat waktu mungkin disebabkan kemampuan yang dimiliki mahasiswa, atau motivasinya. Kemampuan di sini termasuk kemampuan inteligensi dan finansialnya. Masalah ekonomi merupakan alasan utama tidak diselesaikannya pendidikan, selain menikah dini, masalah keamanan, dan masalah sosial lainnya (Latif, Choudhary, & Hammayun, 2015). Karena Universitas Pamulang menerapkan biaya pendidikan yang murah, maka banyak calon mahasiswa yang ingin mencoba-coba kuliah di perguruan tinggi.

Penyebab keluarnya mahasiswa tanpa menyelesaikan pendidikannya, atau tidak dapat menyelesaikan pendidikannya tepat waktu belum dapat dideteksi dengan sistem seleksi saat ini. Padahal mendeteksi mahasiswa yang berisiko tidak menyelesaikan pendidikan pada tahap awal sangat penting (Nurhayati, Kusrini, & Luthfi, 2015), karena dapat digunakan untuk mengambil tindakan untuk mencegah terjadinya *dropout* yang merupakan tantangan terbesar bagi lembaga pendidikan (Pal, 2012). Fenomena *dropout* tidak hanya di Universitas Pamulang, tetapi juga terjadi pada semua tingkat pendidikan di negara manapun, termasuk yang memiliki sosial ekonomi maju (Siri, 2015).

Penggunaan *data mining* untuk menganalisa data mahasiswa dapat menghasilkan penemuan-penemuan baru yang tidak dapat diamati dengan pendekatan statistik tradisional (Sherrill, Eberle, & Talbert, 2011, p. 56). Oleh karena itu diperlukan sistem seleksi penerimaan mahasiswa baru yang dapat memprediksi terjadinya masalah tersebut, sehingga dapat dikurangi.

Teknik *data mining* dan *machine learning* dapat digunakan untuk memprediksi berdasarkan data-data masa lalu. Data mining adalah proses untuk menemukan pola yang berguna dan kecenderungan di dalam kumpulan data yang besar (Larose & Larose, 2015, p. 4). Dari sumber lain, data mining adalah ilmu yang mempelajari tentang pengumpulan, pembersihan, pengolahan, analisis, dan memperoleh wawasan yang berguna dari data (Aggarwal, 2015, p. 1). Salah satu metode yang dapat digunakan untuk memprediksi adalah klasifikasi. Tugas dari klasifikasi adalah memprediksi keluaran variabel/class yang bernilai kategorikal atau polinomial (Kotu & Deshpande, 2015, p. 9). Metode ini digunakan untuk memprediksi calon mahasiswa ke dalam kelompok lulus tepat waktu atau tidak tepat waktu berdasarkan data-data ketika mendaftarkan diri.

Banyak metode data mining yang dapat diterapkan untuk klasifikasi. Algoritma klasifikasi yang populer adalah Decision Trees, Neural Networks, k-Nearest Neighbours, Naive Bayes, dan algoritma Genetik (Yukselturk, Ozekes, & Türel, 2014, p. 119).

Maka pada penelitian ini akan diterapkan beberapa algoritma klasifikasi untuk mencari metode terbaik dalam mengkalsifikasikan data calon mahasiswa di

Jurnal Teknologi 10 (1) pp 25-36 © 2018

Universitas Pamulang. Diharapkan didapat metode klasifikasi terbaik yang dapat digunakan untuk memprediksi ketepatan waktu lulus calon mahasiswa.

Penelitian Terkait

Penelitian terkait kinerja mahasiswa telah banyak dilakukan dan dipublikasikan. Kinerja mahasiswa yang diteliti menyangkut prediksi ketepatan waktu lulus, prediksi *dropout*, dan lain-lain. Sebelum melakukan penelitian, perlu dilakukan kajian terhadap penelitian sebelumnya, agar dapat mengetahui metode, data, maupun model yang sudah pernah digunakan. Kajian penelitian sebelumnya ditujukan untuk mengetahui *state of the art* tentang penelitian prediksi *dropout*, dan prediksi ketepatan waktu lulus menggunakan *data mining*.

Pada penelitian yang dilakukan oleh Pal (Pal, 2012, pp. 35-39) dinyatakan bahwa salah satu tantangan terbesar yang dihadapi lembaga pendidikan adalah mengurangi jumlah peserta didik yang putus studi. Jumlah peserta didik yang putus studi menjadi indikasi kinerja akademik dan manajemen pendaftaran. Hal ini menyebabkan lembaga lebih berfokus pada kekuatan siswa daripada kualitas pendidikan. Pada penelitian ini diterapkan aplikasi data mining untuk menghasilkan model prediktif untuk pengelolaan mahasiswa putus studi, sehingga dapat diketahui mana mahasiswa yang perlu mendapatkan dukungan lebih. Hasil penelitian menunjukkan bahwa algoritma mesin pembelajaran mampu membangun model prediksi secara efektif dari data siswa putus sekolah yang ada.

Pada penelitian yang dilakukan oleh Rai, Saini, dan Jain (Rai, Saini, & Jain, 2014, pp. 142-149) dinyatakan bahwa prediksi awal mahasiswa putus sekolah adalah tugas yang menantang dalam pendidikan tinggi. Analisis data adalah salah satu cara untuk menurunkan tingkat siswa putus sekolah dan meningkatkan angka partisipasi mahasiswa di universitas. Ini adalah kenyataan bahwa mahasiswa putus sekolah cukup sering pada tahun pertama. *Dropout* di universitas disebabkan oleh akademik, keluarga dan alasan pribadi, lingkungan kampus dan infrastruktur universitas dan bervariasi tergantung pada sistem pendidikan yang diterapkan oleh universitas. Pada penelitian ini diusulkan model klasifikasi menggunakan algoritma dan aturan decision tree induction untuk memprediksi apakah mahasiswa akan lulus atau tidak berdasarkan data historis. Hasil menunjukkan bahwa algoritma ID3 adalah pengklasifikasi terbaik dengan akurasi 98%. Hasil penelitian ini juga dapat mengidentifikasi mahasiswa mana yang membutuhkan perhatian khusus untuk mengurangi jumlah *drop-out*.

Pada penelitian yang dilakukan oleh Al-Barrak dan Al-Razgan (Al-Barrak & Al-Razgan, 2015) dinyatakan bahwa kinerja dalam program akademik merupakan salah satu faktor yang paling penting yang mempengaruhi kualitas pendidikan yang lebih tinggi tersedia untuk siswa. Pada karya ilmiah ini, digunakan teknik data mining, khususnya klasifikasi, untuk menganalisis nilai mahasiswa dalam tugas evaluatif yang berbeda untuk mata kuliah data terstruktur. Untuk tujuan ini, dibandingkan tiga pengklasifikasi yang berbeda menggunakan data real dari King Saud University untuk memprediksi kinerja mahasiswa. Di sini diterapkan teknik klasifikasi untuk kedua atribut numerik dan dikategorikan. Hasil kami menunjukkan bahwa model berdasarkan algoritma Naïve Bayes memberikan prediksi yang paling akurat. Selain itu didapatkan model dengan akurasi 91% untuk memprediksi kegagalan mahasiswa dalam mata kuliah.

Pada penelitian yang dilakukan oleh Abu-Oda dan El-Halees (Abu-Oda & El-Halees, 2015, pp. 15-27) dinyatakan bahwa salah satu tantangan terbesar yang dihadapi pendidikan tinggi saat ini adalah memprediksi jalur akademik siswa. Banyak sistem pendidikan tinggi tidak mampu mendeteksi populasi siswa yang cenderung putus karena kurangnya metode intelijen untuk menggunakan informasi, dan bimbingan dari sistem universitas. Untuk mengklasifikasikan dan memprediksi siswa putus sekolah, diusulkan dua pengklasifikasi berbeda, yaitu *Decision Tree* (DT), dan *Naive Bayes* (NB), dan dilatih menggunakan dataset yang telah kumpulkan. Kemudian diuji menggunakan *10-fold cross validation*. Hasilnya menunjukkan bahwa akurasi dari DT mencapai 98,14%, sedangkan NB mencapai 96,86%.

METODE PENELITIAN

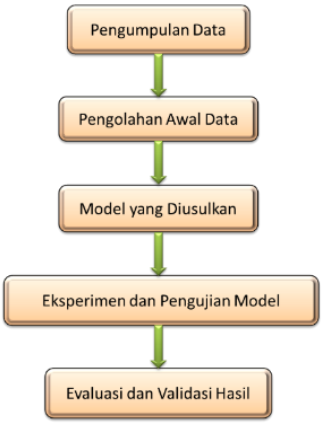
Penelitian ini menggunakan pendekatan kuantitatif. Penelitian kuantitatif bertujuan untuk mencapai pemahaman tentang

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

bagaimana sesuatu dikonstruksi, bagaimana dibangun, atau bagaimana cara kerjanya (Berndtsson, Hansson, Olsson, & Lundell, 2008, p. 13). Penelitian kuantitatif umumnya didorong oleh hipotesis, yang dirumuskan dan diuji secara ketat, dengan tujuan menunjukkan bahwa hipotesisnya salah. Oleh karena itu, salah satu upayanya adalah untuk menyalahkan hipotesis, dan jika hipotesis tahan uji, maka akan dianggap benar setelah terbukti sebaliknya. Aspek kuantitatif adalah untuk menekankan bahwa pengukuran merupakan dasarnya karena memberikan hubungan antara observasi dan formalisasi model, teori, dan hipotesis. Hasil dari penelitian kuantitatif adalah mengembangkan model, teori, dan hipotesis yang berkaitan dengan fenomena alam.

Metode yang digunakan pada penelitian ini adalah eksperimen. Penelitian eksperimen mencakup investigasi hubungan sebab-akibat menggunakan pengujian yang dikontrol sendiri (Dawson, 2009, p. 26). Cukup sering penelitian semieksperimental mendapatkan kendala pada tidak cukupnya akses terhadap sampel, masalah etika dan sebagainya. Eksperimen biasanya dilakukan dalam pengembangan, evaluasi dan pemecahan masalah proyek.

Penelitian ini bertujuan untuk mendapatkan model prediksi ketepatan waktu lulus calon mahasiswa. Karena penelitian yang diakui/diterima harus mengikuti aturan yang diakui (Dawson, 2009, p. 20), maka pada penelitian ini dilakukan dengan mengikuti tahapan seperti Gambar 3.1.



Gambar 1. Tahapan Penelitian

Penjelasan tahapan pada Gambar 1 sebagai berikut ini:

a. Pengumpulan data
Pengumpulan data bukan hanya sekedar mengambil data yang ada, tetapi harus mampu mendeskripsikan data yang ada, dan memiliki kontribusi terhadap pengetahuan. Data tersebut harus dapat memberikan penjelasan, hubungan, perbandingan, prediksi, generalisasi, dan teori (Dawson, 2009, p. 18). Berdasarkan sumbernya, data dibedakan menjadi dua, yaitu:
- Data primer, yaitu data yang dikumpulkan langsung dari sumber data. Pengumpulan data ini memerlukan waktu, dan biaya yang lebih banyak dari data sekunder. Contoh sumber data primer adalah kuisioner, observasi, wawancara, dan eksperimen yang dilakukan langsung oleh peneliti.
- Data sekunder, yaitu data yang diperoleh dari peneliti/pihak lain, walaupun data tersebut sebelumnya digunakan dengan tujuan yang berbeda. Data sekunder dapat diperoleh relatif lebih cepat, dan dengan biaya rendah. Contoh sumber data sekunder adalah kantor statistik baik pemerintah maupun swasta, perpustakaan, toko buku, maupun internet.

Pada penelitian ini digunakan data primer. Data dikumpulkan dari berkas-berkas mahasiswa yang telah lulus yang diperoleh dari tata usaha program studi Teknik Informatika Universitas Pamulang.

b. Pengolahan awal data
Data yang sudah dikumpulkan diolah untuk mengurangi data yang tidak relevan, atau data dengan atribut yang hilang. Pengolahan juga berupa konversi nilai-nilai redundan (berlebihan), atau nilai yang terlalu beragam ke dalam kelompok yang lebih kecil untuk mempermudah pembentukan model.

c. Model/metode yang diusulkan
Untuk menggambarkan alur model/metode yang diusulkan dan menjelaskan cara kerja model/metode yang diusulkan. Model/metode ini

Jurnal Teknologi 10 (1) pp 25-36 © 2018

digambarkan secara skematik dan disertai dengan formula penghitungan. Model/metode yang diusulkan akan dibentuk dari data yang sudah diolah, dan hasil pengolahan model akan diukur dengan model yang ada saat ini.

d. Eksperimen dan pengujian model
Menjabarkan bagaimana eksperimen yang dilakukan hingga terbentuknya model, serta menjelaskan cara menguji model yang terbentuk.

e. Evaluasi dan validasi hasil
Evaluasi dilakukan dengan mengamati hasil prediksi menggunakan algoritma data mining. Validasi digunakan untuk memastikan bahwa hasilnya akan sama ketika dilakukan secara independen. Pengukuran kinerja dilakukan dengan membandingkan nilai ketepatan prediksi dari masing-masing algoritma sehingga dapat diketahui algoritma yang lebih akurat.

Pengumpulan Data

Data yang digunakan adalah data primer, yaitu data yang didapatkan langsung dari objek penelitian dengan melalui hasil pengamatan lapangan dan wawancara. Data-data yang dikumpulkan adalah data-data alumni dan mahasiswa yang masa studinya lebih dari 4 tahun (8 semester) dari Program Studi Teknik Informatika Universitas Pamulang. Data dikumpulkan dari berkas-berkas mahasiswa yang telah lulus yang diperoleh dari tata usaha program studi Teknik Informatika Universitas Pamulang. Spesifikasi dataset kelulusan mahasiswa yang telah dikumpulkan ditunjukkan pada Tabel 1.

Tabel 1. Spesifikasi Dataset Kelulusan Mahasiswa

No	Nama Attribut	Keterangan
1	JenisKelamin	Jenis Kelamin, yaitu Laki-laki, dan Perempuan
2	Jurusan	Jurusan ketika di sekolah, misalnya IPA, IPS, Otomotif, Listrik, dll
3	Matematika	Nilai Matematika pada ujian nasional
4	Inggris	Nilai Bahasa Inggris pada ujian nasional
5	Indonesia	Nilai Bahasa Indonesia pada ujian nasional
6	Status	Status ketepatan waktu kelulusan, yaitu Tepat waktu, dan Tidak tepat

Spesifikasi Dataset yang Digunakan

Penelitian ini dilaksanakan dengan cara menganalisa model yang diusulkan dengan menerapkan pada dataset ketepatan waktu kelulusan mahasiswa. Spesifikasi dan atribut dataset hasil dari tranformasi dataset ketepatan waktu kelulusan mahasiswa ditunjukkan pada Tabel 2.

Tabel 2. Spesifikasi dan Atribut Dataset yang Digunakan

No	Nama Attribut	Keterangan	Jumlah
1	JenisKelamin	Jenis Kelamin, yaitu Laki-laki, dan Perempuan	L=144, P=27
2	Jurusan	Jurusan ketika di sekolah, misalnya IPA, IPS, Otomotif, Listrik, dll	ips=42; listrik=4; tkj=3; ipa=44; otomotif=19; multimedia=11; admin=9; penjualan=8; akuntansi=6; mesin=6; elektro=3; industri=5; rpl=3; sekretaris=1; perhotelan=2; pertanian=2; pariwisata=1; bahasa=1; agama islam=1
	SelisihWaktu	Selisih tahun lulus dengan tahun pendaftaran	Range=0-12 Rata-rata=1,737 SD=2,248
3	Matematika	Nilai Matematika pada ujian nasional	Range=2,0-9,75 Rata-rata=7,526 SD=1,27
4	Inggris	Nilai Bahasa Inggris pada ujian nasional	Range=3,8-10 Rata-rata=7,344 SD=1,27
5	Indonesia	Nilai Bahasa Indonesia pada ujian nasional	Range=4,0-9,8 Rata-rata=7,075 SD=0,991
6	Status	Status	Tepat waktu =

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

	ketepatan waktu kelulusan, yaitu Tepat waktu, dan Tidak tepat	64 Tidak Tepat = 107

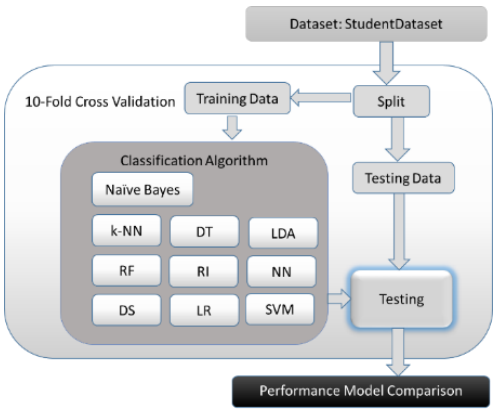
Perancangan Penelitian

Pada penelitian ini diusulkan model prediksi ketepatan waktu lulusan mahasiswa menggunakan beberapa teknik data mining, yaitu 10 algoritma klasifikasi. Keseluruhan akurasi pada pengujian dataset umumnya digunakan untuk mengevaluasi kinerja pengklasifikasi (Zhang & Wang, 2011).

Cross validaton adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran (*learning algorithms*) dengan membagi data menjadi dua segmen, satu segmen digunakan untuk belajar atau data latih, dan yang lain digunakan untuk memvalidasi model (Refaeilzadeh, Tang, & Liu, 2009, p. 532). Dalam *cross validation* kumpulan pelatihan dan validasi harus *crossover* berturut-turut sehingga setiap data memiliki kesempatan tervalidasi.

K-fold cross validation adalah teknik umum untuk memperkirakan kinerja pengklasifikasi. *K-fold cross validation* dilakukan dengan menggunakan kembali dataset yang sama, sehingga menghasilkan k perpecahan dari kumpulan data menjadi *non-overlapping* dengan proporsi pelatihan (k-1)/k dan 1/k untuk pengujian (Korb & Nicholson, 2011, p. 213).

Untuk menguji model yang diusulkan digunakan teknik validasi *10-fold cross validation*. Kerangka kerja penelitian ini ditunjukkan pada Gambar 2.



Gambar 2. Kerangka Kerja Penelitian

Dataset yang telah dikumpulkan (StudentDataset) pertama akan dipecah (*split*) menjadi data latih (*training data*) dan data uji (*testing data*) menggunakan algoritma *10-fold cross validation*. Data latih digunakan untuk melatih algoritma klasifikasi, kemudian data uji digunakan untuk menguji algoritma/model yang telah dilatih. Pada proses pengujian digunakan confusion matrix untuk menghasilkan ukuran kinerja algoritma/model yang diusulkan. Setelah semua algoritma/model diuji, kinerjanya dibandingkan untuk mengetahui model mana yang terbaik.

Teknik Analisis

Sistem intelijen dan model matematis untuk pengambilan keputusan dapat mencapai hasil yang akurat dan efektif hanya jika data yang digunakan dapat diandalkan (Vercellis, 2009, p. 94). Teknik untuk menganalisa algoritma/model yang terbaik dilakukan dengan membandingkan kinerjanya. Pengukuran kinerja model dilakukan dengan dilakukan menggunakan *confusion matrix*. *Confusion matrix* diperoleh dari proses validasi menggunakan *10-fold cross validation*, sehingga model yang terbentuk dapat langsung diuji dengan melakukan 10 kali pengujian.

HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan eksperimen dengan menggunakan komputer untuk melakukan proses penghitungan terhadap model yang diusulkan. Spesifikasi perangkat keras dan sistem operasi yang digunakan pada penelitian ini adalah menggunakan sebuah laptop DELL Inspiron 1440 dengan prosesor Pentium® Dual-Core CPU T4500 @ 2.30 GHz, memori (RAM) 4,00 GB, dan menggunakan sistem operasi Windows 10 Pro 32-bit. Sedangkan perangkat lunak yang digunakan untuk menerapkan model yang diusulkan digunakan RapidMiner Studio 6.0.001 starter edition.

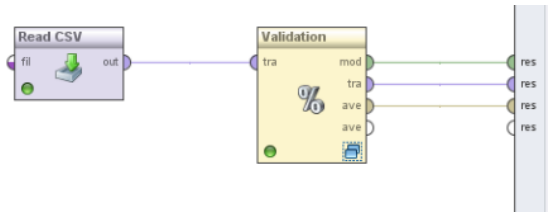
Hasil Pengukuran

Model disimulasikan menggunakan software RapidMiner Studio 6.0.001 starter edition dan dataset mahasiswa yang telah dikumpulkan. Pengukuran kinerja model yang diusulkan ditunjukkan pada subbab selanjutnya.

Jurnal Teknologi 10 (1) pp 25-36 © 2018

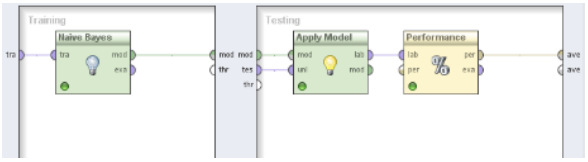
Naïve Bayes (NB)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 3.



Gambar 3. Susunan Validasi Naïve Bayes

Untuk melatih dan menguji model di bagian *training* diisi operator Naïve Bayes, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 4.



Gambar 4. Susunan Operator Naïve Bayes

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 5 didapat akurasinya 60,85%.

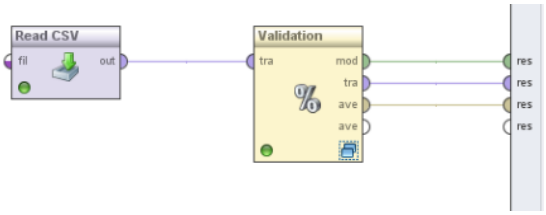
accuracy: 60.85% +/- 14.78% (mikro: 60.82%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	20	23	46.51%
pred. Tidak tepat	44	84	65.62%
class recall	31.25%	78.50%	

Gambar 5. Hasil Pengukuran Naïve Bayes

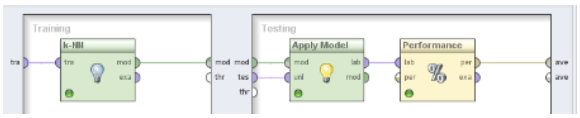
k-Nearest Neighbor (k-NN)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 6.



Gambar 6. Susunan Validasi k-Nearest Neighbor

Untuk melatih dan menguji model di bagian *training* diisi operator *k-Nearest Neighbor*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 7.



Gambar 7. Susunan Operator k-Nearest Neighbor

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 8 didapat akurasinya 57,25%.

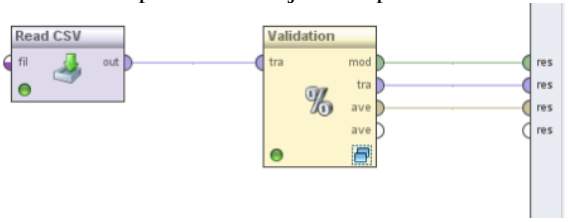
accuracy: 57.25% +/- 13.03% (mikro: 57.31%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	26	35	42.62%
pred. Tidak tepat	38	72	65.45%
class recall	40.62%	67.29%	

Gambar 8. Hasil Pengukuran k-Nearest Neighbor

Random Forest (RF)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 9.

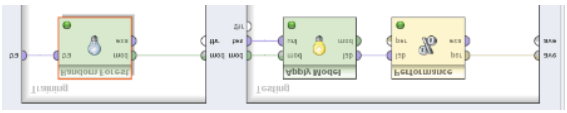


Gambar 9. Susunan Validasi Random Forest

Untuk melatih dan menguji model di bagian *training* diisi operator *Random Forest*, pada *testing* diisi operator *Apply Model* dan

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

Performance. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 10.



Gambar 10. Susunan Operator Random Forest

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 11. didapat akurasinya 62,65%.

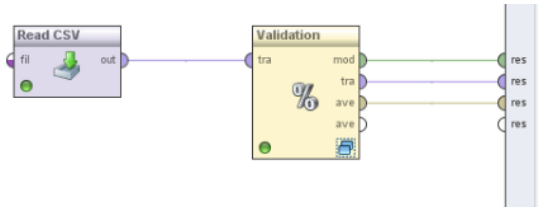
accuracy: 62.65% +/- 31.13% (mikro: 62.57%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	0	0	0.00%
pred. Tidak tepat	64	107	62.57%
class recall	0.00%	100.00%	

Gambar 11. Hasil Pengukuran Random Forest

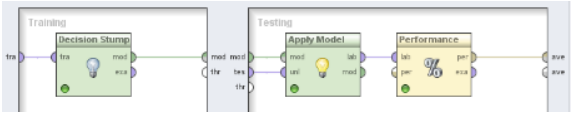
Decision Stump (DS)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 12.



Gambar 12. Susunan Validasi Decision Stump

Untuk melatih dan menguji model di bagian *training* diisi operator *Decision Stump*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 13.



Gambar 13. Susunan Operator Decision Stump

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 14 didapat akurasinya 62,65%.

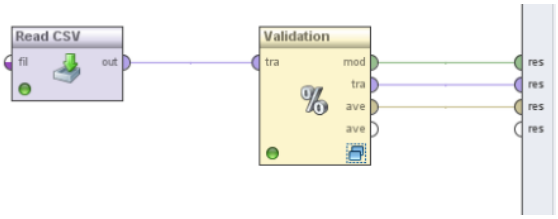
accuracy: 62.65% +/- 30.68% (mikro: 62.57%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	1	1	50.00%
pred. Tidak tepat	63	106	62.72%
class recall	1.56%	99.07%	

Gambar 14. Hasil Pengukuran Decision Stump

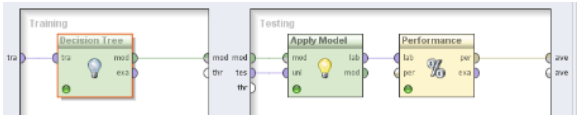
Decision Tree (DT)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 15.



Gambar 15. Susunan Validasi Decision Tree

Untuk melatih dan menguji model di bagian *training* diisi operator *Decision Tree*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 16.



Gambar 16. Susunan Operator Decision Tree

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 17 didapat akurasinya 62,65%.

accuracy: 62.65% +/- 31.13% (mikro: 62.57%)

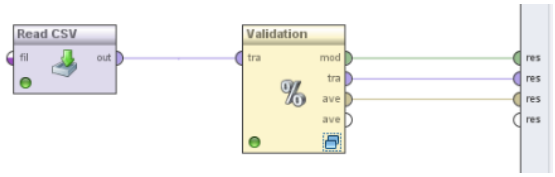
	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	0	0	0.00%
pred. Tidak tepat	64	107	62.57%
class recall	0.00%	100.00%	

Gambar 17. Hasil Pengukuran Decision Tree

Rule Induction (RI)

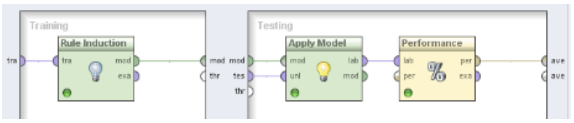
Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 18.

Jurnal Teknologi 10 (1) pp 25-36 © 2018



Gambar 18. Susunan Validasi Rule Induction

Untuk melatih dan menguji model di bagian *training* diisi operator *Rule Induction*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 19.



Gambar 19. Susunan Operator Rule Induction

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 20 didapat akurasinya 62,03%.

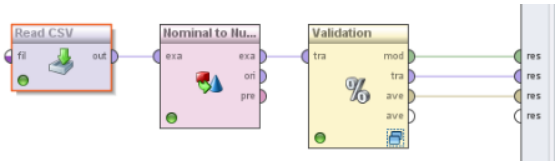
accuracy: 62.03% +/- 24.49% (mikro: 61.99%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	12	13	48.00%
pred. Tidak tepat	52	94	64.38%
class recall	18.75%	87.85%	

Gambar 20. Hasil Pengukuran Rule Induction

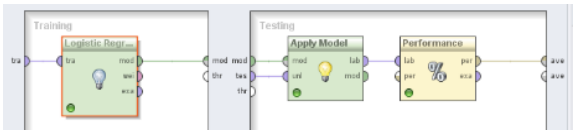
Linear Regression (LR)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Karena Linear Regression tidak mendukung masukan nominal, maka dikonversi menggunakan Nominal to Numeric. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 21.



Gambar 21. Susunan Validasi Linear Regression

Untuk melatih dan menguji model di bagian *training* diisi operator *Linear Regression*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 22.



Gambar 22. Susunan Operator Linear Regression

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 23 didapat akurasinya 57,91%.

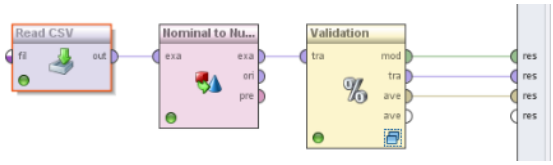
accuracy: 57.91% +/- 8.93% (mikro: 57.89%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	26	34	43.33%
pred. Tidak tepat	39	73	65.77%
class recall	40.62%	68.22%	

Gambar 23. Hasil Pengukuran Linear Regression

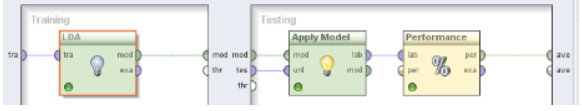
Linear Discriminant Analysis (LDA)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Karena Linear Regression tidak mendukung masukan nominal, maka dikonversi menggunakan Nominal to Numeric. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 24.



Gambar 24. Susunan Validasi Linear Discriminant Analysis

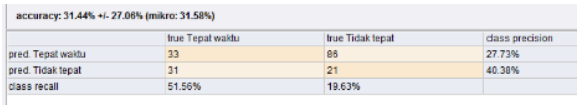
Untuk melatih dan menguji model di bagian *training* diisi operator *Linear Discriminant Analysis*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 25.



Gambar 25. Susunan Operator Linear Discriminant Analysis

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 26 didapat akurasinya 31,44%.

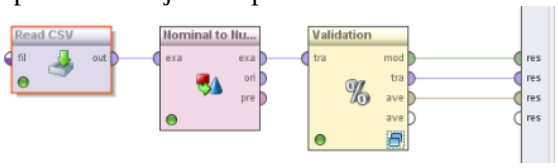

accuracy: 31.44% +/- 27.08% (mikro: 31.58%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	33	86	27.73%
pred. Tidak tepat	31	21	40.38%
class recall	51.56%	19.63%	

Gambar 26. Hasil Pengukuran Linear Discriminant Analysis

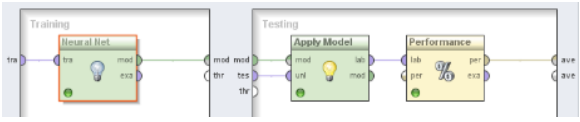
Neural Network (NN)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Karena Linear Regression tidak mendukung masukan nominal, maka dikonversi menggunakan Nominal to Numeric. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 27.



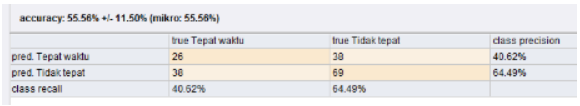
Gambar 27. Susunan Validasi Neural Network

Untuk melatih dan menguji model di bagian *training* diisi operator *Neural Network*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 28.



Gambar 28. Susunan Operator Neural Network

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 29 didapat akurasinya 55,56%.

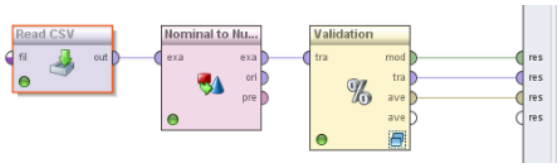

accuracy: 55.56% +/- 11.50% (mikro: 55.56%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	26	38	40.62%
pred. Tidak tepat	36	69	64.49%
class recall	40.62%	64.49%	

Gambar 29. Hasil Pengukuran Neural Network

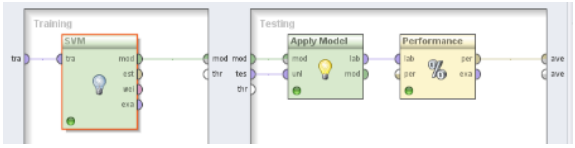
Support Vector Machine (SVM)

Dataset yang disimpan dalam format CSV (*Comma Separated Value*) dibuka menggunakan operator Read CSV. Karena Linear Regression tidak mendukung masukan nominal, maka dikonversi menggunakan Nominal to Numeric. Kemudian keluarannya dihubungkan ke *Validation* (*X-Validation*) dengan *10-fold cross validation*. Susunan operator ditunjukkan pada Gambar 30.



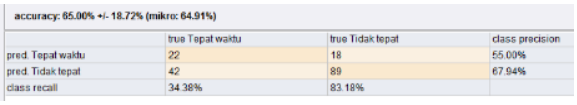
Gambar 30. Susunan Validasi Support Vector Machine

Untuk melatih dan menguji model di bagian *training* diisi operator *Support Vector Machine*, pada *testing* diisi operator *Apply Model* dan *Performance*. Susunan operator *training* dan *testing* ditunjukkan pada Gambar 31.



Gambar 31. Susunan Operator Support Vector Machine

Kemudian model dieksekusi sehingga didapat hasil kinerja model seperti pada Gambar 32 didapat akurasinya 65,00%.


accuracy: 65.00% +/- 18.72% (mikro: 64.91%)

	true Tepat waktu	true Tidak tepat	class precision
pred. Tepat waktu	22	18	55.00%
pred. Tidak tepat	42	89	67.94%
class recall	34.38%	83.18%	

Gambar 32. Hasil Pengukuran Support Vector Machine

Pembahasan

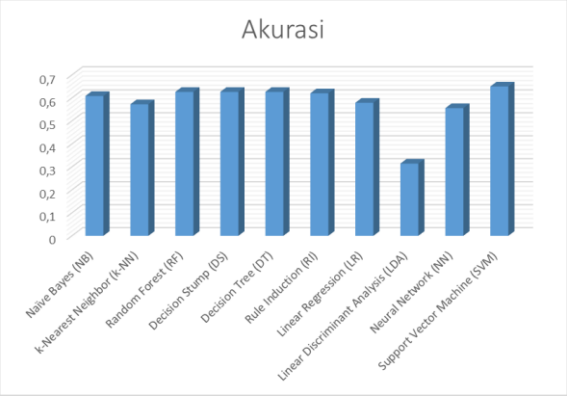
Dari pengujian algoritma/model yang diusulkan diperoleh ukuran akurasi seperti ditunjukkan pada Tabel 3.

Jurnal Teknologi 10 (1) pp 25-36 © 2018

Tabel 3. Hasil Pengukuran Akurasi Algoritma/Model

No	Algoritma/model	Akurasi
1	Naïve Bayes (NB)	60,85%
2	k-Nearest Neighbor (k-NN)	57,25%
3	Random Forest (RF)	62,65%
4	Decision Stump (DS)	62,65%
5	Decision Tree (DT)	62,65%
6	Rule Induction (RI)	62,03%
7	Linear Regression (LR)	57,91%
8	Linear Discriminant Analysis (LDA)	31,44%
9	Neural Network (NN)	55,56%
10	Support Vector Machine (SVM)	65,00%

Hasil pengukuran yang didapat divisualisasikan dalam grafik batang 3D seperti pada Gambar 33.



Gambar 33. Visualisasi Akurasi Algoritma/Model

Dari hasil pengukuran diperoleh model terbaik yaitu *Support Vector Machine* (SVM) dengan akurasi 65.00%. Tetapi akurasi ini masih jauh dari nilai *excellent* (sangat baik).

KESIMPULAN

Berdasarkan hasil implementasi dan pengukuran algoritma/model yang diusulkan diperoleh algoritma/model terbaik, yaitu *Support Vector Machine* (SVM) dengan akurasi 65.00%. Tetapi akurasi ini masih jauh dari nilai *excellent* (sangat baik).

DAFTAR PUSTAKA

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data Mining in Higher Education: University Student Dropout Case Study. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(1), 15-27. doi:10.5121/ijdkp.2015.5102

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Switzerland: Springer International Publishing.

Al-Barrak, M. A., & Al-Razgan, M. S. (2015). Predicting Students' Performance Through Classification: A Case Study. *Journal of Theoretical and Applied Information Technology*, 167-175.

Berndtsson, M., Hansson, J., Olsson, B., & Lundell, B. (2008). *Thesis Projects: A Guide for Students in Computer Science and Information Systems* (2nd ed.). London: Springer-Verlag.

Bisri, A., & Wahono, R. S. (2015). Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree. *Journal of Intelligent Systems*, *1*(1), 27-32.

Dawson, C. W. (2009). *Projects in Computing and Information Systems A Student's Guide* (2nd ed.). Great Britain: Pearson Education.

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian Artificial Intelligence* (2nd ed.). Florida: CRC Press.

Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner*. Massachusetts: Elsevier Inc.

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2nd ed.). New Jersey: John Wiley & Sons, Inc.

Latif, A., Choudhary, A. I., & Hammayun, A. A. (2015). Economic Effects of Student Dropouts: A Comparative Study. *Journal of Global Economics*, *3*(2), 1-4. doi:10.4172/2375-4389.1000137

Manhães, L. M., Cruz, S. M., & Zimbrão, G. (2014). Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining. *Data Mining for Educational Assessment and Feedback (ASSESS 2014)* (pp. 1-7). New York: Aspiring Minds.

Nurhayati, S., Kusrini, & Luthfi, E. T. (2015). Prediksi Mahasiswa Drop Out menggunakan Metode Support Vector Machine. *Jurnal Ilmiah SISFOTENIKA*, *5*(1), 82-93.

Pal, S. (2012). Mining Educational Data Using Classification to Decrease Dropout Rate of Students. *International journal of multidisciplinary sciences and engineering*, *3*(5), 35-39.

Rai, S., Saini, P., & Jain, A. K. (2014). Model for Prediction of Dropout Student Using ID3 Decision Tree Algorithm. *International*

ISSN : 2085 – 1669
e-ISSN : 2460 – 0288

Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), *2*(1), 142-149.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. Liu, & M. T. Özsu, *Encyclopedia of Database Systems* (pp. 532-538). Arizona: Springer US.

Sherrill, B., Eberle, W., & Talbert, D. (2011). Analysis of Student Data for Retention Using Data Mining Techniques. *7th Annual National Symposium on Student Retention* (pp. 65-66). Charleston: C-IDEA.

Siri, A. (2015). Predicting Students' Dropout at University Using Artificial Neural Networks. *Italian Journal of Sociology of Education*, *7*(2), 225-247. doi:10.14658/pupj-ijse-2015-2-9

Vercellis, C. (2009). *Business Intelligence- Data Mining and Optimization for Decision Making.* West Sussex: John Wiley & Sons.

Wahyudin, N. (2015). Analisis Faktor-Faktor yang Mempengaruhi Keunggulan Bersaing untuk Meningkatkan Kinerja Perguruan Tinggi Swasta (PTS) pada Sekolah Tinggi dan Akademi di Semarang. *Holistic Journal of Management Research*, *3*(2), 77-92.

Yasmiati, Wahyudi, & Susilo, A. (2017). Pengembangan Aplikasi Data Mining dengan Algoritma C4.5 dan Apriori di Fakultas Teknologi Informatika Universitas Respati Indonesia. *Jurnal Teknologi*, *9*(1), 31-41.

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and e-Learning*, *17*(1), 118-133. doi:10.2478/eurodl-2014-0008

Zhang, H., & Wang, Z. (2011). A Normal Distribution-Based Over-Sampling Approach to Imbalanced Data Classification. *Advanced Data Mining and Applications - 7th International Conference* (pp. 83-96). Beijing: Springer.