Jurnal Tekno Insentif | ISSN (p): 1907-4964 | ISSN (e): 2655-089X | Vol. 13 | No. 1 | Halaman 16-25
DOI: https://doi.org/10.36787/jti.v13i1.78 April 2019

PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4.5, NAÏVE BAYES, KNN, DAN SVM

Sri Widaningsih
Universitas Suryakancana
sriwida@unsur.ac.id

Abstrak - Kelulusan mahasiswa merupakan salah satu hal yang harus diperhatikan karena termasuk ke dalam Standar Penjaminan Mutu Internal suatu perguruan tinggi. Program Studi Teknik Informatika merupakan salah satu prodi yang ada di Universitas Suryakancana. Untuk kelulusan terdapat standar yang akan dicapai oleh prodi yaitu waktu studi yaitu empat tahun dan IPK minimal 3,00. Untuk dapat mencapai mutu lulusan tersebut dibutuhkan suatu prediksi tingkat kelulusan dengan standar yang telah ditetapkan untuk mahasiswa yang masih menjalankan studi sehingga dapat dilakukan langkah antisipasi dari awal sehingga dapat menanggulangi terjadinya permasalahan dalam bidang akademik. Untuk memprediksi tingkat kelulusan dan IPK standar tersebut digunakan metode *data mining* dengan fungsi klasifikasi. Metodologi penerapan data *mining* ini menggunakan tahapan *Discovery Knowledge of Databases* (KDD) dimulai dari tahap *selecting, preprocessing, transformation, data mining* dan *evaluation/interpretation.* Teknik yang akan digunakan untuk model data mining klasifikasi ini terdiri dari empat algoritma yaitu C4.5, *Support vector machine* (SVM), *k-nearest neigbor* (kNN,) dan *Naïve Bayes*. Metoda klasifikasi terdiri dari variabel-variabel prediktor dan satu variabel target. Variabel-variabel prediktor terdiri dari jenis kelamin dan indeks prestasi dari semester 3 hingga 6. Perangkat lunak yang digunakan untuk mengolah data yaitu *software Rapid Miner*. Hasil akhir dari keempat algoritma tersebut diperoleh bahwa algoritma *Naïve Bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu dan IPK ≥ 3 dengan nilai *accuracy (*76,79%), *error* (23,17%) , dan AUC (0,850).

Kata kunci : kelulusan, data *mining*, klasifikasi, *naïve bayes,* kNN, SVM, C4.5

Abstract *- Graduate students is one factor that must be considered because it is included in the Quality Assurance Standards. The Informatic Engineering Department is one of the department at Suryakancana University. For the graduate standard that will be approved by the department, time period graduation is four years and a minimum GPA of 3.00. To achieve the required level of success, the predetermined standard level must be set for students to be anticipated from the start so they can overcome problems in the academic field. To predict the graduation rate and GPA standard the data mining method is used with the classification function. This data mining implementation methodology uses the stages of Discovery Knowledge of Databases (KDD) starting from selecting, preprocessing, transformation, data mining and evaluation / interpretation. The technique that will be used for this data mining model consists of four algorithms, such as C4.5, Support vector machine (SVM), the nearest k-neigbor (kNN,) and Naïve Bayes. The classification method consists of predictor variables and one target variable. Predictor variables consist of gender and achievement index from semester 3 to 6. The software used to process data is RapidMiner software. The final results of the following four algorithms generated from the Naïve Bayes algorithm are the best algorithms for predicting timely student graduation and GPA ≥ 3 with accuracy (76.79%), errors (23.17%), and AUC (0.850).*

Keywords : *graduation, data mining, classification, naïve bayes, kNN, SVM, C4.5*

1. PENDAHULUAN

Kelulusan mahasiswa merupakan salah satu bidang yang termasuk ke dalam Standar Penjaminan Mutu Internal (SPMI) suatu perguruan tinggi. Salah satu standar yang ditetapkan oleh program studi Teknik Informatika Universitas Suryakancana ditetapkan untuk lulusan teknik informatika menghasilkan lulusan tepat waktu yaitu maksimal delapan semester dengan IPK minimal 3,00. Saat ini prodi Teknik Informatika telah menghasilkan sekitar 500 orang lulusan dari tahun 2006 hingga 2016. Setiap tahun jumlah mahasiswa mengalami peningkatan. Saat ini data-data yang tersimpan di dalam *database* belum digali lebih dalam untuk mendapatkan suatu informasi atau pengetahuan yang dapat digunakan lebih lanjut untuk pengambilan suatu keputusan. Dengan semakin meningkatnya jumlah mahasiswa dan standar IPK lulusan yang semakin tinggi juga, maka diperlukan suatu penelitian mengenai prediksi nilai dan waktu lulusan mahasiswa saat ini berdasarkan pada data-data masa lalu. Dari data mahasiswa yang ada di *database* tentunya terdapat beberapa variabel-variabel prediktor yang dapat digunakan untuk memprediksi variabel target yaitu waktu dan nilai IPK lulusan. Salah satu metode untuk memprediksi yaitu menggunakan *data mining* , dimana akan dicari pola yang terdapat pada *database* lulusan untuk memprediksi waktu kelulusan empat tahun dengan nilai IPK minimal 3, 00. Teknik-teknik yang akan digunakan termasuk ke

dalam fungsi klasifikasi dengan beberapa algoritma yaitu C4.5, Naïve Bayes, k- *Nearest Neighbour* (kNN) dan *Support Vector Machine* (SVM).

Rumusan masalah dalam penelitian ini yaitu bagaimana algoritma C4.5, Naïve Bayes, kNN, dan SVM dapat membantu memprediksi waktu dan nilai kelulusan mahasiswa Teknik Informatika dan dari keempat teknik tersebut manakah yang memberikan prediksi hasil yang paling baik.

Tujuan dari penelitian ini diantara-Nya yaitu :

1. Memanfaatkan data-data mahasiswa yang tersimpan di dalam *database* untuk memperoleh informasi mengenai kelulusan
2. Memprediksi kelulusan tepat waktu dengan nilai IPK minimal 3,00 menggunakan algoritma dan teknik C4.5, *naïve bayes, kNN* dan SVM.
3. Membandingkan dari keempat teknik dan algoritma yang digunakan yang memberikan hasil paling baik dimana paling mendekati hasil sebenarnya

Sedangkan manfaat dari penelitian ini yaitu :

1. Memberikan gambaran mengenai pola kelulusan yang ada saat ini
2. Sebagai masukan untuk mengantisipasi mahasiswa-mahasiswa yang tidak sesuai dengan target standar lulusan

2. KAJIAN PUSTAKA DAN PERUMUSAN HIPOTESIS

Definisi Data Mining

Data mining merupakan salah satu teknik untuk menggali atau "menambang" pengetahuan dari sekumpulan besar data. Data *mining* merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya yang dapat dipahami dan bermanfaat bagi pemilik data (Larose, 2005). Terdapat beberapa teknik yang digunakan untuk data *mining* seperti yang diungkapkan Turban, et al (2011) *data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Data *mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar.

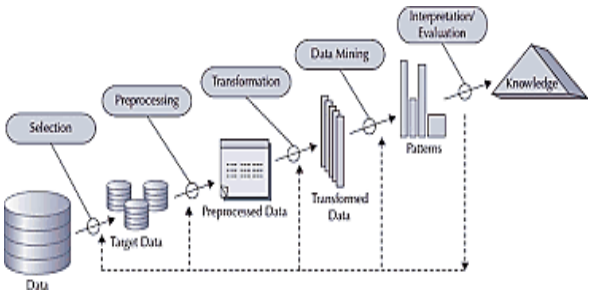
Data *mining* biasanya mengolah data dari *database* dengan ukuran yang besar. Dari data tersebut dilakukan pencarian pola atau *trend* sesuai dengan tujuan dari penerapan data *mining* tersebut. Hasil dari pengolahan data *mining* tersebut selanjutnya dapat digunakan untuk pengambilan keputusan maupun analisis yang dibutuhkan.

Terdapat beberapa alasan mengapa ilmu data *mining* dibutuhkan saat ini diantara-Nya terdapat sejumlah besar data di suatu perusahaan atau organisasi yang hanya tersimpan di dalam *database* tanpa dianalisis lebih lanjut untuk digunakan untuk perkembangan perusahaan atau organisasi tersebut. Selain itu dengan perkembangan internet yang sangat pesat, memberikan dampak positif dengan kemudahan akses data dengan berbagai perangkat *hardware* dan *software* yang memiliki daya komputasi dan kapasitas yang luar biasa. Sedangkan dilihat lingkungan luar, tekanan kompetisi untuk memperluas pangsa pasar dan keuntungan juga semakin meningkat sehingga dibutuhkan cara lain dengan menggali informasi yang tersimpan pada data yang dimiliki perusahaan atau organisasi tersebut.

Meskipun algoritma *data mining* biasanya diterapkan untuk ukuran data yang besar, beberapa algoritma bisa juga diterapkan untuk ukuran data yang relatif kecil. Kumpulan data yang digunakan dalam data *mining* sederhana dalam struktur dimana baris menjelaskan kasus-kasus individu (juga disebut sebagai pengamatan atau contoh) dan kolom menggambarkan atribut atau variabel dari kasus-kasus. Pilihan algoritma yang akan digunakan tergantung pada jenis data (yaitu, nominal, ordinal, rasio atau interval).

Knowledge Discovery in Databases (KDD)

Data mining merupakan salah satu bagian dari proses *Knowledge Discovery in Databases* (KDD). KDD merupakan proses mencari informasi yang lebih bernilai, lebih mudah dipahami dan baru dari penyimpanan data yang besar dan kompleks. Proses KDD menafsirkan hasil yang diperoleh dari sekumpulan data dengan menggabungkan dengan ilmu lainnya. Proses KDD dimulai dengan menetapkan tujuan dan diakhiri dengan evaluasi (Tomar & Agarwal, 2013). Tahapan dari KDD dapat dilihat pada Gambar 1 di bawah ini :



Gambar 1. Tahapan proses KDD

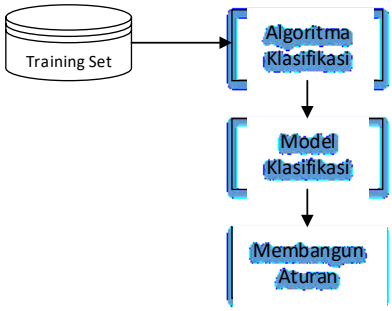
Klasifikasi

Algoritma data *mining* dapat dibagi menjadi tiga (Neelamegam & Ramaraj, 2013), yaitu *supervised,*

unsupervised, dan *semi-supervised*. Dalam *supervised learning*, algoritma bekerja pada sekumpulan data yang telah diberi label atau telah diketahui kelasnya. Pada *supervised learning*, data belum diketahui label atau kelasnya, algoritma digunakan untuk mengelompokkan data berdasarkan kemiripannya. Sedangkan dalam *semi supervised learning*, sebagian kecil data telah memiliki label bersama dengan sejumlah data yang belum memiliki label. Klasifikasi termasuk ke dalam *supervised learning*.

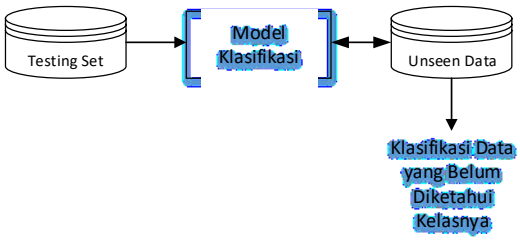
Proses klasifikasi dibagi menjadi dua tahap (Annasaheb & Verma, 2016) yaitu :

1. Tahap membangun model
   Pada langkah ini model klasifikasi dibangun berdasarkan data yang telah ditentukan kelasnya. Data sampel yang digunakan disebut sebagai data pelatihan atau data pembelajaran (*training set*). Proses ini disebut sebagai proses induksi yang ditunjukkan pada Gambar 2 .



Gambar 2. Tahapan membangun model

2. Tahap menggunakan model klasifikasi
   Pada tahap ini model diterapkan pada data yang belum diketahui kelasnya. Proses penerapan model klasifikasi untuk memprediksikan kelas label dari data dalam himpunan menggunakan data uji (*testing set*), proses ini disebut deduksi. Proses ini dapat dilihat pada Gambar. 3.



Gambar 3. Tahapan menggunakan model

Algoritma Klasifikasi

Dalam klasifikasi, terdapat variabel target yang bersifat kategoris yang dibagi menjadi kelas yang telah ditentukan, seperti kelas nasabah yang bermasalah atau tidak, hewan yang masuk ke dalam klasifikasi reptil, amfibi, mamalia, burung, atau ikan.

Model *data mining* memeriksa sejumlah besar data, setiap *record* berisi informasi tentang variabel target serta sekumpulan *input* atau prediktor variabel. Setiap algoritma klasifikasi yang digunakan akan menghasilkan model yang paling sesuai menghubungkan antara data *input* dan kelas klasifikasi yang telah diketahui sebelumnya.

Setiap algoritma bisa menghasilkan klasifikasi yang berbeda. Algoritma terbaik dapat dilihat dari data yang diklasifikasikan secara benar oleh model dengan data sebenarnya atau seberapa akurat model dapat memprediksi kelas klasifikasi.

Terdapat beberapa algoritma dan teknik yang digunakan pada data *mining*, diantara-Nya C4.5, *support vector machine*, k-*nearest neighbor* ,Naïve Bayes , dan *artificial neural network* (Nikam, 2015). Berikut ini penjelasan mengenai C4.5, SVM, Naïve Bayes, dan kNN.

1. Algoritma C4.5

Algoritma C4.5 termasuk ke dalam pohon keputusan. Struktur sebuah pohon keputusan seperti pada *flowchart* , dimana setiap simpul internal (simpul bukan daun) melakukan pengujian pada atribut, masing-masing cabang merupakan sekumpulan hasil, dan masing-masing simpul daun (atau simpul terminal) menjadi label kelas. Simpul paling atas dalam pohon adalah simpul akar. Algoritma pohon keputusan merupakan *supervised learning*, maka memerlukan pra klasifikasi variabel sasaran. Sekumpulan data *training* harus dipersiapkan untuk membentuk algoritma dengan nilai-nilai variabel target. Dalam metode ini dievaluasi semua atribut menggunakan ukuran statistik berupa *information gain* dan perhitungan entropi.

Information gain merupakan perolehan informasi atau ukuran efektivitas suatu atribut dalam mengklasifikasikan data . Persamaan (1) merupakan rumus *information gain* yaitu (Suyatno, 2017).

$$Gain\ (S,A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

Dimana :

A : atribut

$|S_v|$: jumlah sampel untuk nilai v

$|S|$: jumlah seluruh sampel data

Entropi adalah keberagaman suatu data. Pada persamaan (2) merupakan rumus entropi :

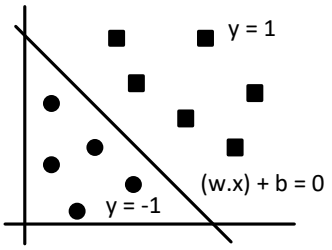
$$Entropy\ (S) = -\sum p_i log_2 p_i \quad .....(2)$$

Dimana :

p_i = porsi atau rasio antara jumlah sampel kelas i dengan jumlah semua sampel pada himpunan data

2. *Support Vector Machine*

Pada SVM dapat mengklasifikasikan data linier dan non linier. Data *input* merupakan nilai variabel-variabel prediktor sedangkan *output* merupakan variabel target dimana saling bergantung. Dalam tujuan SVM adalah untuk menemukan fungsi klasifikasi terbaik untuk membedakan antara anggota dari dua kelas dalam data pelatihan. Metrik untuk konsep fungsi klasifikasi "terbaik" dapat diwujudkan secara geometris. Untuk *dataset* terpisah secara linear, fungsi klasifikasi linier berhubungan dengan *hyperplane* pemisah f (x) yang melewati tengah dua kelas, memisahkan keduanya (Neelamegam & Ramaraj, 2013). Bidang pemisah linier dalam SVM dapat dilihat pada Gambar 4 berikut .



Gambar 4. Bidang pemisah linier

3. Naïve Bayes

Klasifikasi Bayes merupakan klasifikasi secara statistik, klasifikasi ini dapat memprediksi peluang keanggotaan kelas seperti probabilitas suatu tupel merupakan milik kelas tertentu. Klasifikasi ini berdasarkan teorama Bayes. Pada Naïve Bayes berasumsi bahwa efek dari nilai atribut pada kelas tertentu independen dari nilai-nilai dari atribut lainnya

Rumus Naive Bayes (Suyatno, 2017). Rumus algoritma naive Bayes ditunjukkan pada persamaan (3) berikut :

$$P(Y \mid X) = P(Y) \prod P(X_i \mid Y) \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Dimana :

P(X | Y) :probabilitas data dengan vektor X pada kelas Y

P(Y) : probabilitas awal kelas Y dan P(Xi | Y) adalah probabiltas independen kelas Y pada semua fitur dalam vektor X

4. k Nearest Neighbor

Algoritma ini merupakan pendekatan mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama. Disebut juga dengan pembelajar yang malas (*lazy learners*) karena hanya melihat kedekatan dengan tetangga (*neihbor*). Pada persamaan (4) merupakan salah satu rumus jarak yang digunakan dalam K Nearest Neighbor adalah jarak *Euclidian* (Gorunescu, 2011) :

$$d_{Euclidian\,(x,y)} = \sqrt{\sum_i (x_i - y_i)^2} \quad \ldots\ldots\ldots (4)$$

A. Alat Evaluasi

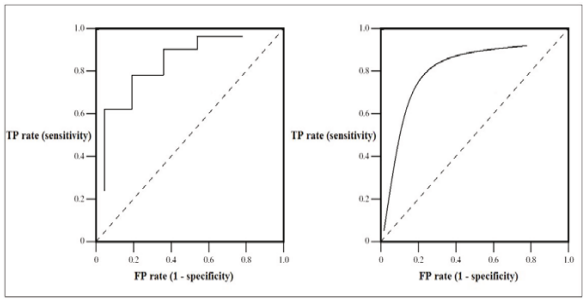
Klasifikasi biner merupakan model statistik dan perhitungan yang membagi kumpulan data menjadi dua kelompok yaitu positif dan negatif. Tabel 1 berikut adalah *confusion matrix* untuk menjelaskan ukuran performansi klasifikasi.

Tabel 1. *Confusion Matrix*

Aktual	Prediksi	
	Positive	*Negative*
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

Ukuran performansi termasuk ke dalam tahapan evaluasi. Beberapa ukuran performansi untuk teknik klasifikasi yaitu akurasi, *error*, dan *Area Under Receiver Operating Characteristics* (ROC) c*urve* (AUC). Akurasi adalah suatu ukuran rasio prediksi yang benar terhadap total jumlah sampel dievaluasi. Error adalah ukuran rasio prediksi yang salah terhadap total jumlah sampel yang dievaluasi. AUC (*the area under curve*) dihitung untuk mengukur perbedaan performansi.

Kurva ROC menunjukkan akurasi dan me mbandingkan klasifikasi secara visual dengan *false positive rate (specificity)* sebagai garis horizontal dan *true positive rate (sensitivity)* sebagai garis vertikal. seakan-akan menggambarkan tawar menawar antara sensitivitas (*benefit*) dan 1- spesifisitas (*cost*) , yang tujuannya adalah untuk menentukan *cut off point* pada uji *diagnostic* yang bersifat kontinu. Kurva ROC dapat dilihat pada Gambar 5.



Gambar 5. Kurva ROC untuk kasus diskrit dan kontinu

Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = Klasifikasi sangat baik
b. 0.80-0.90 = Klasifikasi baik
c. 0.70-0.80 = Klasifikasi cukup
d. 0.60-0.70 = Klasifikasi buruk
e. 0.50-0.60 = Klasifikasi salah

$$Akurasi = \frac{TP + TN}{TP + TN + FN + FP} \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$Error = \frac{FP + FN}{TP + TN + FN + FP} \dots\dots\dots\dots\dots\dots\dots (6)$$

B. RapidMiner

Dalam pengolahan data *mining* umumnya digunakan *software* sebagai alat bantu. Beberapa *software data mining* diantaramya RapidMiner, weka, clementine, tanagra dan lain-lain. Menurut www.rapidminer.com, software rapidminer digunakan untuk merancang aliran secara visual untuk menganalisis *data science* dan *machine learning* di dalam tim mulai dari analis hingga pakar.

Rapidminer memiliki kemudahan dalam penggunaan, dapat mengumpulkan data dari semua sumber seperti basis data, *cloud*, dokumen, media sosial dan aplikasi bisnis. Selain itu dapat mengeksplorasi dan memvisualisasi data secara statistik. Tersedia beberapa model mesin pembelajaran dan model validasi.

Penelitian Terdahulu

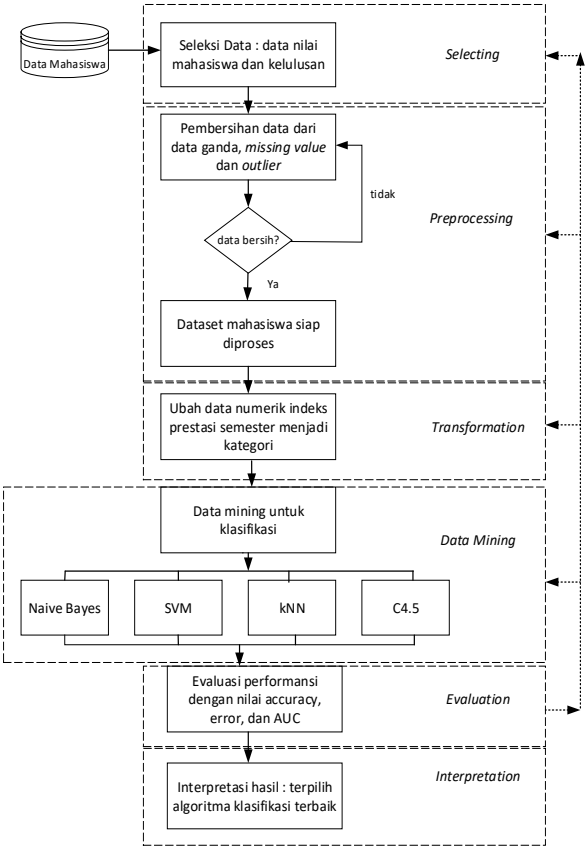
Beberapa penelitian mengenai penerapan data *mining* untuk menggali informasi yang terdapat pada *database* mahasiswa diantara-Nya yaitu oleh Sillueta (2016) yang memprediksi kelulusan dengan algoritma kNN dimana diperoleh akurasi rata-rata sekitar 70%. Subna dan Muhardi (2016) memprediksikan hasil akademik dengan menggunakan variabel prediktor dosen, motivasi, kedisiplinan, ekonomi dan hasil belajar dengan menggunakan C4.5. Prediksi waktu kelulusan juga dilakukan dengan algoritma *artificial* neural *network* (Meinanda, dkk, 2009). Priati (2016) melakukan analisis perbandingan pada tiga buah algoritma yaitu C4.5, Naïve Bayes dan CART untuk memprediksi kelulusan mahasiswa, dan diperoleh bahwa algoritma C4.5 memberikan nilai akurasi yang tertinggi, tetapi nilai AUC tertinggi adalah algoritma Naïve Bayes.

3. METODE PENELITIAN

Model yang digunakan dalam penelitian ini mengikuti tahapan yang ada pada proses *knowledge discovery in database* (KDD). Pada Gambar 6 adalah penjelasan untuk masing-masing tahap yang akan dilakukan .

Berikut ini penjelasan tiap tahap penelitian pada Gambar 6.

1. *Selection*
   Pada tahap ini dilakukan seleksi data lulusan yang terdiri dari variabel-variabel prediktor dan satu target variabel. Variabel target yaitu klasifikasi lulusan yang lulus tepat waktu yaitu 4 tahun atau kurang dan memiliki nilai IPK minimal 3,00. Sedangkan variabel-variabel prediktor yaitu , jenis kelamin, indeks prestasi semester 3, 4,5 dan 6.
2. *Preprocessing*
   Data yang diambil sesuai dengan banyaknya lulusan di prodi Teknik Informatika. Dari data yang diambil dilakukan pembersihan data apabila terdapat data yang hilang, data ganda atau bersifat *outlier.*



Gambar 6. Tahapan penelitian berdasarkan KDD

3. *Transformation*
   Setelah data bersih dari kesalahan , selanjutnya dilakukan transformasi pada data sesuai dengan jenis data pada tahapan transformasi dimana jenis akan dikelompokkan menjadi data yang bersifat kategori. Berikut pada Tabel 2 adalah kategori untuk variabel prediktor dan variabel target .

Tabel 2. Kategori Variabel Prediktor dan Variabel Target

Variabel Target	Kategori	
Kelulusan dan IPK	Sesuai (lulus ≤ 4 tahun dan IPK ≥ 3,00) Tidak Sesuai	
Variabel Prediktor		
Jenis Kelamin	Perempuan Laki-laki	
Indeks Prestasi Semester (IPS) 3-6	Besar	IPS ≥ 3,00
	Sedang	2,75 ≤ IPS < 3,00
	Kecil	IPS < 2,75

4. *Data Mining*
   Pada tahap ini dilakukan pemilihan teknik *data mining* yang sesuai. Untuk fungsi klasifikasi digunakan algoritma C4.5, SVM, kNN dan Naïve Bayes. Karena klasifikasi merupakan *supervised learning* maka berikut ini adalah tahapan dalam model *supervised learning* (Larose ,2005).
5. *Evaluation*
   Tahap ini digunakan untuk mengevaluasi hasil-hasil prediksi yang dihasilkan oleh keempat algoritma dan dipilih metode algoritma yang menghasilkan nilai mendekati klasifikasi data sebenarnya. Evaluasi dilakukan dengan menggunakan metode *Confusion Matrix* dan kurva ROC (*Receiver Operating Characteristic*). Nilai performansi yang digunakan yaitu *accuracy* dan *error*.

4. ANALISIS DAN PERANCANGAN

Data Selection

Sumber data mentah yang digunakan dalam penelitian ini adalah data mahasiswa Teknik Informatika pada tahun 2008 hingga tahun 2013. Data nilai berasal dari transkrip nilai mahasiswa yang terdapat pada *database* bagian akademik Fakultas Teknik. Sedangkan data kelulusan berasal dari data wisuda mahasiswa setiap tahunnya.

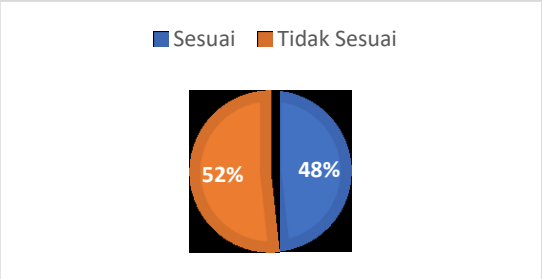
Preprocessing Data

Untuk mengetahui IPK setiap semesternya dilakukan perhitungan kembali dari transkrip nilai karena terdapat beberapa redudansi data nilai karena faktor mengulang kembali mata kuliah. Nilai yang diambil merupakan nilai tertinggi yang didapat. Contoh dari pengumpulan data disusun pada Tabel 3 di bawah ini

Tabel 3. Data awal mahasiswa teknik Informatika yang belum ditransformasi

NO	NPM	JK	IPS-3	IPS-4	IPS-5	IPS-6	IPK	LULUS	Tepat	>3	SESUAI
1	103015520108001	L	3.60	3.58	3.05	3.70	3.49	2012	Y	Y	Y
2	103015520108002	L	3.60	3.68	3.30	3.85	3.55	2012	Y	Y	Y
3	103015520108003	L	3.05	3.05	3.30	3.40	3.14	2012	Y	Y	Y
4	103015520108006	L						TL			
5	103015520108007	L	3.30	3.42	3.55	3.70	3.36	2012	Y	Y	Y
6	103015520108008	P	3.05	3.21	3.20	3.40	3.21	2012	Y	Y	Y
7	103015520108010	L						TL			
8	103015520108011	L						TL			
9	103015520108012	L	3.35	3.00	2.85	3.10	3.14	2012	Y	Y	Y
10	103015520108014	L	2.80	2.68	2.90	3.30	2.79	2014	T	T	T
11	103015520108015	L						TL			
12	103015520108016	L						TL			
13	103015520108017	L	3.30	2.79	3.45	3.55	3.19	2012	Y	Y	Y
14	103015520108020	L						TL			
15	103015520108021	L	3.05	2.79	3.00	2.85	2.92	2013	T	T	T
16	103015520108022	P						TL			
17	103015520108023	P	3.00	2.95	3.45	4.00		TL			
18	103015520108024	L	3.05	2.84	3.25	3.25	3.08	2013	T	Y	T
19	103015520108025	L	2.55	2.53	2.60	2.70	2.60	2013	T	T	T
20	103015520108027	L						TL			

Berdasarkan data kelulusan mahasiswa Teknik informatika angkatan tahun 2008 hingga 2016 yang lulus tepat waktu dengan IPK ≥ 3,00 sebanyak 52% dan diberi klasifikasi kelas "sesuai", dapat dilihat pada Gambar 7 berikut.



Gambar 7. Perbandingan klasifikasi kelas "sesuai" dengan "tidak sesuai"

Transformation

Dataset yang akan diolah dimasukkan pada tabel data awal yang nantinya akan dilakukan transformasi pada beberapa jenis data yang bersifat numerik yaitu nilai indeks prestasi semester 3-6 (IPS 3-6) . Proses transformasi nilai IPS 3-6 berdasarkan pada pembagian yang terdapat di Tabel 2. Karena data NPM tidak masuk sebagai variabel prediktor, maka data tersebut tidak akan diproses pada pengolahan data. Bentuk *dataset* yang telah ditransformasi ada pada Tabel 4 di bawah ini.

Tabel 4. Data yang Telah Ditransformasi

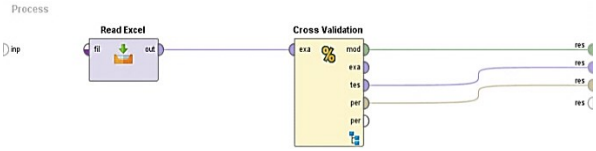
No	JK	ips-3	ips-4	ips-5	ips-6	Sesuai
1	p	besar	besar	sedang	besar	Y
2	l	sedang	sedang	sedang	sedang	Y
3	l	sedang	sedang	besar	besar	Y
4	p	sedang	sedang	kecil	sedang	Y
5	l	sedang	sedang	sedang	sedang	Y
6	l	sedang	kecil	sedang	sedang	T
7	l	sedang	sedang	sedang	besar	Y
8	p	sedang	sedang	sedang	sedang	T
9	l	besar	sedang	kecil	besar	T
10	l	kecil	kecil	kecil	kecil	T

Data Mining

Pada pengolahan data, dilakukan tahap pemodelan untuk proses klasifikasi yaitu menerapkan algoritma Naïve Bayes , SVM, kNN dan C4.5 (Gambar 8). Pengolahan data menggunakan perangkat lunak *rapid miner* 8.0.001. Jumlah data yang akan diolah sebanyak 466 data. Proses pengambilan data pada *Rapidminer* untuk algoritma Naïve Bayes, kNN dan C4.5 langsung pada format *excel* dari data yang telah ditransformasi seperti pada Gambar 8 di bawah . Selanjutnya dilakukan *cross validation* untuk data yang telah diambil. Teknik validasi yang digunakan untuk pada proses klasifikasi adalah *k-Fold Cross Validation.*

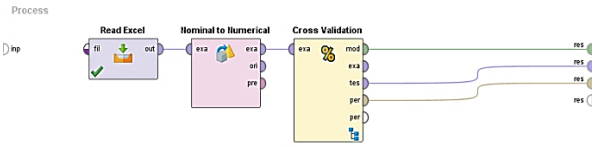
K-Fold Cross-validation adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja

model klasifikasi dimana data dipisahkan menjadi dua bagian yaitu data proses latih (*training*) dan data uji . *k-Fold Cross Validation* digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Nilai k diambil 10 *fold* sehingga dari 466 data akan menjadi 10 *subset* data dengan ukuran sama yaitu sekitar 46,6 atau 47 data. Dari masing-masing 10 *subset* tersebut, 419 data menjadi data latih dan 47 data menjadi data uji.



Gambar 8. Proses pengambilan data untuk algoritma naïve bayes, kNN, dan C4.5

Untuk Algoritma SVM dilakukan pengubahan data dari bentuk nominal menjadi bentuk numerik karena algoritma SVM pada *Rapidminer* hanya dapat diproses jika data berbentuk numerik. Pengubahan data menggunakan fungsi *Nominal to Numerical* seperti pada Gambar 9 di bawah ini.

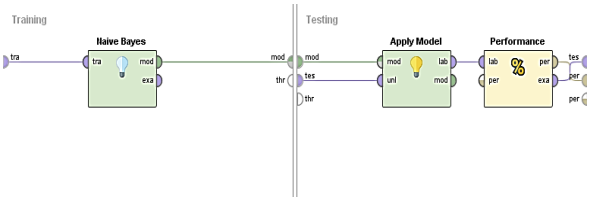


Gambar 9. Proses pengambilan data untuk algoritma SVM

Pada proses validasi dilakukan penerapan algoritma-algoritma data *mining* sesuai dengan yang digunakan pada penelitian ini yaitu algoritma Naïve Bayes, kNN, SVM , dan C4.5.

A. Hasil Perhitungan Klasifikasi dengan RapidMiner

Pada data *training* diterapkan algoritma Naïve Bayes untuk teknik klasifikasi seperti pada Gambar 10 di bawah ini



Gambar 10. Model klasifikasi dengan algoritma naïve bayes

Output hasil performansi dari algoritma ini berupa pengklasifikasian mahasiswa-mahasiswa yang termasuk ke dalam klasifikasi kelas "sesuai" atau "tidak sesuai" yang diterapkan pada data *testing*. Jumlah data yang diprediksi dengan benar oleh algoritma Naïve Bayes ditunjukkan dalam tabel 5 *confusion matrix* berikut .

Tabel 5. *Confusion Matrix* Algoritma Naïve Bayes

Aktual	Prediksi	
	YA	TIDAK
YA	162	60
TIDAK	48	196

Keterangan tabel 5 adalah:

a. Jumlah data sebenarnya yang SESUAI dan diprediksi SESUAI adalah 162.
b. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi TIDAK SESUAI adalah 196.
c. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi SESUAI adalah 48.
d. Jumlah data sebenarnya yang SESUAI dan diprediksi TIDAK SESUAI adalah 60.

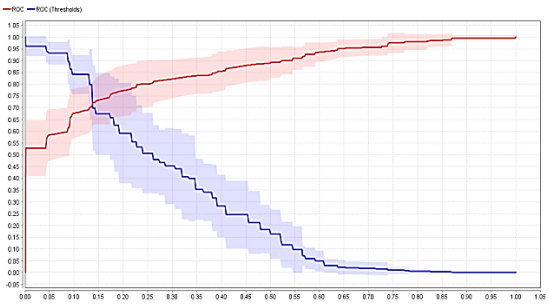
Evaluasi untuk model ini menggunakan nilai akurasi , *error* , dan *Area Under Curve* (AUC) ROC (*Receiver Operating Characteristic*). Grafik ROC untuk algoritma naïve bayes ditunjukkan pada Gambar 11.

Akurasi dari model yaitu :

$$Akurasi = \frac{162 + 196}{162 + 60 + 48 + 196} = 76{,}79\%$$

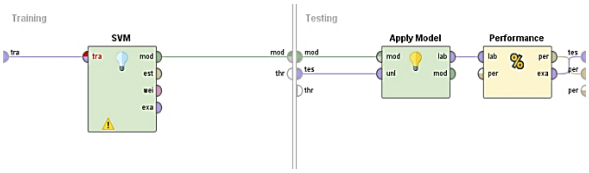
$$Error = \frac{48 + 60}{162 + 60 + 48 + 196} = 23{,}17\%$$

AUC = 0,850



Gambar 11. Grafik ROC algoritma naïve bayes

Pada data *training* diterapkan algoritma SVM untuk teknik klasifikasi seperti pada Gambar 12 di bawah ini.



Gambar 12. Model klasifikasi dengan algoritma SVM

Jumlah data yang diprediksi dengan benar oleh algoritma SVM ditunjukkan dalam Tabel 6 *confusion matrix* berikut .

Tabel 6. *Confusion Matrix* Algoritma SVM

Aktual	Prediksi	
	YA	TIDAK
YA	199	23
TIDAK	98	146

Keterangan tabel 6 adalah:

a. Jumlah data sebenarnya yang SESUAI dan diprediksi SESUAI adalah 199.
b. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi TIDAK SESUAI adalah 146.
c. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi SESUAI adalah 98.
d. Jumlah data sebenarnya yang SESUAI dan diprediksi TIDAK SESUAI adalah 23.

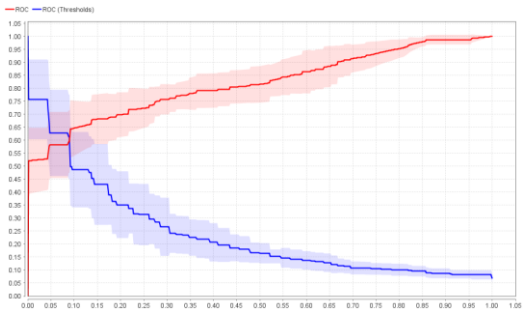
Akurasi dari model yaitu :

$$Akurasi = \frac{199 + 146}{199 + 23 + 98 + 146} = 74{,}04\%$$

$$Error = \frac{98 + 23}{199 + 23 + 98 + 146} = 25{,}96\%$$

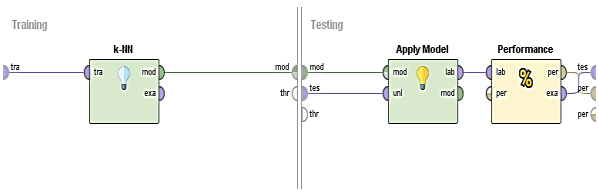
AUC = 0,797

Grafik ROC untuk algoritma SVM ditunjukkan pada Gambar 13.



Gambar 13. Grafik ROC algoritma SVM

Pada data *training* diterapkan algoritma kNN untuk teknik klasifikasi seperti pada Gambar 14 di bawah ini. Nilai k yang diambil sebanyak 3 karena memberikan hasil performansi yang lebih baik dibandingkan dengan k = 1, 2 atau lebih dari 3



Gambar 14. Model klasifikasi dengan algoritma kNN

Jumlah data yang diprediksi dengan benar oleh algoritma kNN ditunjukkan dalam *confusion matrix* seperti yang ditunjukkan pada Tabel 7 berikut .

Tabel 7. *Confusion Matrix* Algoritama kNN dengan k = 3

Aktual	Prediksi	
	YA	TIDAK
YA	161	61
TIDAK	88	156

Keterangan tabel 7 adalah:

a. Jumlah data sebenarnya yang SESUAI dan diprediksi SESUAI adalah 161.
b. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi TIDAK SESUAI adalah 156.
c. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi SESUAI adalah 88.
d. Jumlah data sebenarnya yang SESUAI dan diprediksi TIDAK SESUAI adalah 61.

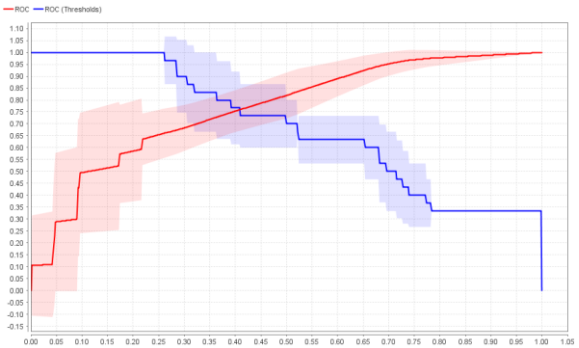
Akurasi dari model yaitu :

$$Akurasi = \frac{161 + 156}{161 + 61 + 88 + 156} = 68{,}05\%$$

$$Error = \frac{61 + 88}{161 + 61 + 88 + 156} = 31{,}97\%$$

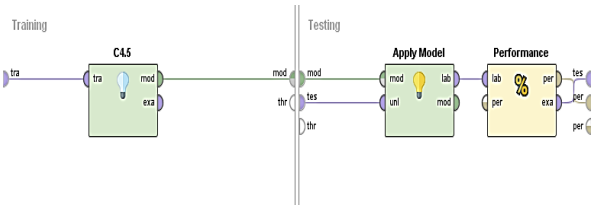
AUC = 0,725

Grafik ROC untuk algoritma kNN ditunjukkan pada Gambar 15.



Gambar 15. Grafik ROC algoritma kNN

Pada data *training* diterapkan algoritma kNN untuk teknik klasifikasi seperti pada Gambar 16 di bawah ini.



Gambar 16. Model klasifikasi dengan algoritma C4.5

Jumlah data yang diprediksi dengan benar oleh algoritma C4.5 ditunjukkan dalam tabel 8 *confusion matrix* berikut .

Tabel 8. *Confusion Matrix* Algoritama C4.5

Aktual	Prediksi	
	YA	TIDAK
YA	159	63
TIDAK	49	195

Keterangan tabel 8 adalah:

a. Jumlah data sebenarnya yang SESUAI dan diprediksi SESUAI adalah 159.
b. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi TIDAK SESUAI adalah 195.
c. Jumlah data sebenarnya yang TIDAK SESUAI dan diprediksi SESUAI adalah 49.
d. Jumlah data sebenarnya yang SESUAI dan diprediksi TIDAK SESUAI adalah 63.

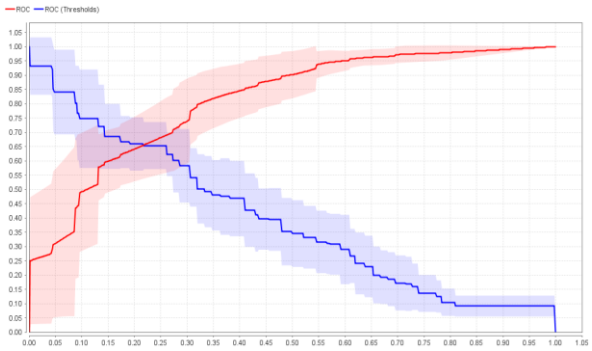
Akurasi dari model yaitu :

$$Akurasi = \frac{159 + 195}{159 + 195 + 63 + 49} = 75{,}96\%$$

$$Error = \frac{63 + 49}{159 + 195 + 63 + 49} = 24{,}03\%$$

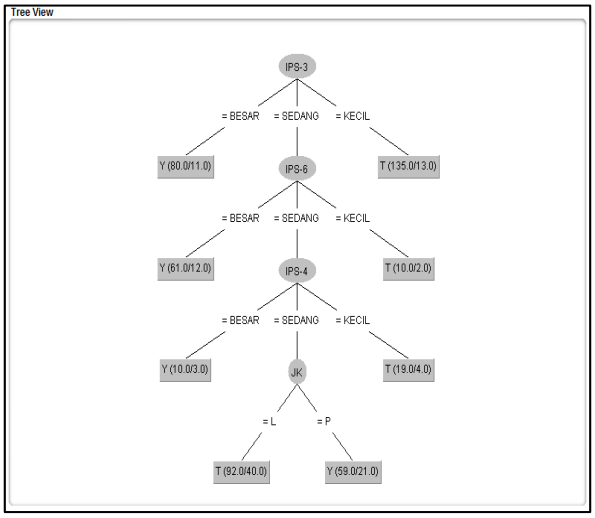
AUC = 0,811

Grafik ROC untuk algoritma C4.5 ditunjukkan pada Gambar 17.



Gambar 17. Grafik ROC algoritma C4.5

Gambar 18 di bawah merupakan pohon keputusan yang dihasilkan oleh algoritma C4.5 . Dari lima variabel yang digunakan terlihat empat variabel yang membentuk pohon, tiga variabel merupakan variabel Indeks Prestasi Semester (IPS) dan Jenis Kelamin (JK). Indeks prestasi yang membentuk pohon yaitu IPS3, IPS4, dan IPS6. Sedangkan variabel IPS5, tidak ada dalam pohon keputusan. Yang menjadi simpul akar adalah IPS3 karena memiliki *gain* tertinggi. Jika IPS3 besar maka dapat diprediksi mahasiswa lebih banyak yang "SESUAI " sedangkan jika IPS3 kecil diprediksi mahasiswa tersebut lebih banyak "TIDAK SESUAI" , tetapi jika nilai IPS3 sedang, maka lihat IPS6 (IP Semester 6). Jika IPS6 besar maka diprediksi mahasiswa lebih banyak "SESUAI" dan jika kecil maka diprediksi mahasiswa lebih banyak "tidak sesuai" sedangkan jika IPS6 sedang maka lihat IPS4. Jika IPS4 besar maka diprediksi mahasiswa lebih banyak "SESUAI", jika kecil maka diprediksi mahasiswa lebih banyak "TIDAK SESUAI". Tetapi jika sedang, maka lihat JK (jenis kelamin). Jika laki-laki diprediksi lebih banyak "TIDAK SESUAI", sedangkan jika perempuan diprediksi lebih banyak "SESUAI".



Gambar 18. Pohon keputusan algoritma C4.5

5. KESIMPULAN DAN SARAN

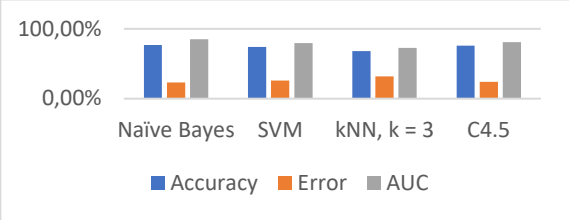
Kesimpulan

Hasil performansi pada setiap model yaitu *accuracy*, *error* dan ROC selanjutnya akan dibandingkan untuk mengetahui algoritma mana yang lebih baik dalam memprediksi kelulusan mahasiswa yang tepat waktu dan IPK ≥ 3. Tabel 9 berikut merupakan tabel perbandingan performansi setiap model.

Tabel 9. Komparasi Nilai Performansi Setiap Algoritma

Algoritma	***Accuracy***	***Error***	**AUC**
Naïve Bayes	76,79%	23,17%	0,850
SVM	74,04%	25,96%	0,797
kNN, k = 3	68,05%	31,97%	0,725
C4.5	75,96%	24,03%	0,811

Dari hasil perbandingan terlihat bahwa algoritma Naïve Bayes memiliki nilai yang paling baik untuk semua kategori performansi dibandingkan dengan algoritma lainnya. Untuk nilai *accuracy* dan AUC nilai terbesar adalah yang terbaik, sedangkan untuk *error* adalah nilai yang terkecil. Nilai AUC untuk Naïve Bayes dan C4.5 termasuk kedalam kategori "baik", sedangkan untuk algoritma SVM dan kNN termasuk kedalam kategori "cukup". Grafik perbandingan keempat algoritma dapat dilihat pada Gambar 19 di bawah ini.



Gambar 19. Grafik perbandingan performansi

Dari hasil penelitian dapat disimpulkan bahwa :

1. Dengan teknik data *mining* dapat diperoleh informasi-informasi yang terdapat di dalam *database* mahasiswa seperti pola kelulusan
2. Dari algoritma-algoritma yang diuji semuanya dapat digunakan untuk memprediksi tingkat kelulusan yang diharapkan, dilihat dari nilai AUC semua algoritma terdapat dalam kategori "baik" dan "cukup"
3. Dari hasil evaluasi diperolah hasil bahwa algoritma naïve bayes yang paling baik untuk memprediksi tingkat kelulusan yang diharapkan karena memiliki nilai akurasi tertinggi dan error terkecil dibandingkan dengan algoritma lainnya.

Saran

Berikut ini adalah beberapa saran yang dapat dilakukan untuk penelitian selanjutnya :

1. Untuk meningkatkan nilai akurasi dari algoritma dapat ditambahkan variabel-variabel lain yang diperkirakan mempengaruhi tingkat kelulusan
2. Dapat dilakukan pengujian algoritma lain selain algoritma yang digunakan dalam penelitian ini

6. DAFTAR PUSTAKA

Annasaheb, A.B., & Verma, V.K. (2016). Classification Techniques: A Recent Survey. International Journal of Emerging Technologies in Engineering Research (IJETER), 4, 51-54.

Gorunescu, Florin. (2011). Data Mining Concept ,Model Technique.Springer-Verlag Berlin Heidelberg

Larose. 2005. Discovering Knowledge in Data.Canada: Wiley-Interscience

Meinanda, dkk. (2009). Prediksi Masa Studi Sarjana dengan Artificial Neural Network, Internetworking Indonesia Journal, 1, 31-35

Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in Data mining: An Overview. International Journal of P2P Network Trends and Technology (IJPTT), 3, 1-5.

Nikam, N. N,.(2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal Of Computer Science & Technology, 8, 13-19

Priati.(2016). Kajian Perbandingan Teknik Klasifikasi Algoritma C4.5, Naïve Bayes Dan CART Untuk Prediksi Kelulusan Mahasiswa (Studi Kasus : STMIK Rosma Karawang).Media Informatika. 15, 1-17

Sillueta, C.Y. (2016). Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Dengan Metode Klasifikasi Dan Algoritma Knearest Neighbor Berbasis Desktop (Studi Kasus : Fakultas Teknologi Informasi, Program Studi Teknik Informatika, Tugas Akhir

Subna, E., & Muhardi. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar, Jurnal CoreIT, 2, 41-44

Suyatno. (2017) Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio- Science and Bio -Technology, 5 , 241-266

Turban, Efraim, Sharda, R. dan Delen, D. (2011). Decision Support Systems and Business Intelligent Systems. Pearson