Faktor Exacta 11 (1): 65-74, 2018
p-ISSN: 1979-276X
e- ISSN: 2502-339X

KOMPARASI ALGORITMA KLASIFIKASI MENENTUKAN KELULUSAN MATA KULIAH PADA UNIVERSITAS

Nahot Frastian
Senna Hendrian
V.H. Valentino
Program Studi Informatika
Fakultas Teknik dan Ilmu Komputer
Universitas Indraprasta PGRI
Jl. Nangka No. 58 C, Tanjung Barat, Jagakarsa, Jakarta Selatan 12530
Email: nahotfrastian@gmail.com, sennahendrian@yahoo.co.id, valentino_na70@yahoo.com

Abstract Graduation of the course is a reflection of the seriousness of individual students every semester of the lecture. Students who attend all the learning process in the classroom or face-to-face in each course will have more chance to graduate than the students who are rarely present, some lecturer's evaluation of the graduation of the course are attendance, task, UTS and UAS. From the above problem the writer tries to get the appropriate algorithm to determine the graduation of each student's course, this research will use the data mining classification technique with 3 methods of classification algorithm such as Algorithm C4.5 (decision tree), Naïve Bayes, and Random Forest with label result pass or not pass. Based on the results of tests conducted using the same dataset on the 3 algorithms through comparison to the value of AUC and Confusion Matrix, obtained the value of Area Under Curve (AUC) of 1,000 from Naïve Bayes model, while the largest Accuracy or Confusion Matrix value is in C4 algorithm .5 (decision tree) with a value of 97.78%. Thus, the proper algorithm for this case is the C4.5 Algorithm (decisiontree).

Keywords: Course graduation, data mining classification, Algorithm C4.5 (decision tree), Naïve Bayes, and Random Forest.

PENDAHULUAN

Lulus kuliah tepat waktu bagi mahasiswa yang sedang mengenyam pendidikan tinggi merupakan sebuah target yang harus dicapai. Hal tersebut bergantung pada mahasiswa itu sendiri. Berbicara masalah lulus kuliah tidak terlepas dari tahapan kelulusan seluruh mata kuliah yang diambil setiap semesternya. Faktor utama penyebab lulus dan tidaknya mata kuliah adalah melalui penilaian dosen mata kuliah terhadap mahasiswanya, bagi mahasiswa yang mengikuti semua kegiatan proses belajar di kelas atau tatap muka setiap mata kuliah akan lebih punya peluang lulus dibanding dengan mahasiswa yang jarang hadir, dasar lulus atau tidaknya mata kuliah ditentukan oleh beberapa kriteria penilaian antara lain: nilai kehadiran, tugas, UTS dan UAS. Untuk menentukan lulus atau tidaknya mata kuliah, dibutuhkan penerapan teknologi yang serba sistematis dan diharapkan lebih baik dari pola sebelumnya. Dalam memperoleh algoritma klasifikasi yang terbaik, maka dibutuhkan beberapa penerapan metode algoritma. Dalam makalah ini metode algoritma klasifikasi yang akan digunakan antara lain *Algoritma C4.5 (decision tree), naïve bayes*, dan *Random Forest*. Begitu banyak teknologi dan metode algoritma yang bisa digunakan untuk mengolah data (*data mining*), namun pada kenyataannya masih banyak dosen yang belum menerapkan teknologi dan metode algoritma, sementara data tersebut dapat memberikan informasi yang penting dan bermanfaat.

Pada penulisan ini, fokus utama *data mining* yang akan dibahas adalah klasifikasi, dimana algoritma yang akan digunakan untuk mengklasifikasikan dataset adalah *Algoritma C4.5 (decision tree), Naïve Bayes*, dan *Random Forest*. Sementara data *training* yang digunakan adalah data penilaian mahasiswa terhadap mata kuliah Pemrograman 2 disalah satu universitas di Jakarta, berikut kriteria penilaian kelulusan mata kuliah antara lain nilai kehadiran, nilai

tugas, nilai UTS dan nilai UAS. Data penilaian terhadap salah satu mata kuliah ini terdiri dari 87 *record* dengan tujuan akhir adalah keputusan lulus atau tidak terhadap mata kuliah tersebut.

Data Mining

Adalah istilah yang diciptakan untuk menggambarkan proses pergeseran melalui database besar untuk mencari pola yang menarik dan sebelumnya tidak diketahui (Lior dan M. Oded, 2015)

Klasifikasi adalah salah satu peran utama dalam *data mining*. Klasifikasi adalah tipe analisis data yang dapat membantu orang menentukan kelas label dari sampel yang ingin di klasifikasi (Lan. et al, 2007).

Klasifikasi

Klasifikasi adalah bagian dari prediksi, dimana nilainya berupa label. Klasifikasi menentukan *class* atau grup untuk masing-masing contoh data, *input* dari klasifikasi adalah atribut dari data *sample*, dan *output*nya adalah *class* dari data *samples* itu sendiri, umumnya dalam *machine learning* untuk membangun model klasifikasi digunakan metode *supervised learning*.

Metode *supervised learning* merupakan cara untuk menemukan hubungan antara atribut masukan dengan atribut target, hubungan yang ditemukan kemudian disebut model. Dalam klasifikasi kita dapat menentukan orang atau objek kedalam suatu kategori tertentu, contoh untuk masalah klasifikasi adalah menentukan apakah mahasiswa "lulus" atau "tidak lulus" terhadap mata kuliah tertentu. Informasi tentang mahasiswa sebelumnya digunakan sebagai bahan untuk melatih algoritma dalam mendapatkan *rule* atau aturan.

Algoritma C4.5

C4.5 adalah algoritma yang mempunyai masukan berupa *training samples* data, contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya dan *samples* yang merupakan *field-field* data yang nantinya akan digunakan sebagai tolak ukur dalam melakukan klasifikasi data.

Algoritma dasar dari C4.5 adalah sebagai berikut:

1. Pohon yang dihasilkan berupa pohon terbalik
2. Pada tahap awal, semua contoh training adalah akar
3. Atribut adalah kategori
4. Contoh di partisi secara berulang berdasarkan atribut yang dipilih
5. Atribut tes dipilih dari data *heuristic* atau pengukuran statistik

Tahapan algoritma C4.5 adalah sebagai berikut:

1. Menyiapkan *data training*
2. Pilih atribut sebagai akar

Untuk memilih atribut akar, sumbernya ada pada nilai *Gain* tertinggi dari atribut yang ada. Dan untuk mendapatkan nilai *Gain*, perlu ditentukan terlebih dahulu nilai *Entropy*.

Rumus *Entropy* :

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

S = Himpunan Kasus
n = Jumlah Partisi *S*
pi = Proporsi dari *Si* terhadap *S*

Rumus *Gain* :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

S = Himpunan Kasus
A = Atribut
n = Jumlah Partisi Atribut
|Si| = Jumlah Kasus pada partisi ke-*i*
|S| = Jumlah Kasus dalam *S*

Decision Tree

Decision Tree atau Pohon Keputusan adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasi. Pada pohon keputusan, masing-masing node internal (*non-leaf*) merepresentasikan sebuah variabel atribut (atribut prediksi atau fitur) dan masing-masing cabang merepresentasikan satu keadaan dari variabel ini. Masing-masing dari tiga daun (*leaf*) menspesifikasikan nilai yang diharapkan dari kelas variabel (variabel yang akan di prediksi). Aspek penting dari prosedur untuk membangun pohon keputusan adalah pemisahan kriteria (*split criterion*) termasuk kriteria untuk membuat cabang dan kriteria terakhir (*stop criterion*), kriteria yang digunakan untuk menghentikan pencabangan.

Pohon keputusan dibuat menggunakan himpunan dari data yang digunakan sebagai data pembelajaran (*training dataset*). Himpunan yang berbeda yang disebut *test dataset* digunakan untuk melakukan pengujian untuk mengecek model.

Pohon keputusan menawarkan banyak keuntungan, antara lain :

1. Fleksibilitas untuk berbagai tugas *data mining*, seperti klasifikasi, regresi, clustering dan seleksi fitur.
2. Cukup jelas dan mudah diikuti (ketika dipadatkan).
3. Fleksibilitas dalam menangani berbagai input data: nominal, numerik dan tekstual.
4. Adaptasi di dataset pengolahan yang mungkin memiliki kesalahan atau nilai-nilai yang hilang.
5. Kinerja prediktif tinggi untuk upaya komputasi yang relatif kecil
6. Tersedia dalam berbagai paket *data mining* melalui berbagai *platform*
7. Berguna untuk dataset besar (dalam kerangka *ensemble*).

Naïve Bayes

Merupakan metode yang tidak memiliki aturan, *naïve bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi setiap klasifikasi pada data training. *naïve bayes* merupakan metode klasifikasi populer dan termasuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes, Simple Bayes* dan *Independence Bayes* (Bramer, 2007).

Klasifikasi *Bayes* di dasarkan pada *teorema bayes*, diambil dari nama seorang ahli matematika yang juga menteri Prebysterian Inggris, Thomas Bayes (1702-1761). Yaitu:

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

Keterangan:

Y : Data dengan kelas yang belum diketahui
X : Hipotesis data y yang merupakan suatu kelas spesifik
P(x|y) : Probabilitas hipotesis x berdasarkan kondisi y (*posteriori probability*)
P(x) : Probabilitas hipotesis x (*prior probability*)
P(y|x) : Probabilitas y berdasarkan kondisi pada hipotesis x
p(y) : Probabilitas dari y

Random Forest

Random Forest adalah pengklasifikasi yang terdiri dari kumpulan pengklasifikasi pohon terstruktur {h(x,Θk), k=1, ...} dimana {Θk} adalah vektor acak terdistribusi yang identik independen dan masing-masing pohon melemparkan unit suara untuk kelas paling populer di input x[5].

Random forest merupakan pengembangan dari *Algoritma C4.5 (decision tree)* dengan menggunakan beberapa *decision tree*, dimana setiap *decision tree* telah dilakukan *training data* menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut *subset* yang bersifat acak. Dan dalam perkembangannya, sejalan dengan bertambahnya *dataset*, maka *tree* pun ikut berkembang. Penempatan *tree* yang saling berjauhan membuat apabila terdapat *tree* disekitar *tree x* berarti pohon tersebut merupakan perkembangan *tree x* [4].

Rapid Miner

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan bahasa java dibawah lisensi GNU *Public License* dan *Rapid Miner* dapat dijalankan disistem operasi manapun. Menggunakan *Rapid Miner* tidak membutuhkan kemampuan koding khusus, karena semua fitur sudah tersediakan. *Rapid Miner* dikhususkan untuk penggunaan data mining.

Evaluasi dan Validasi

Validasi adalah proses mengevaluasi akurasi prediksi dari sebuah model, validasi mengacu untuk mendapatkan prediksi dengan menggunakan model yang ada kemudian membandingkan hasil yang diperoleh dengan hasil yang diketahui (Gorunescu, 2011).

Mengevaluasi akurasi dari model klasifikasi sangat penting, akurasi dari sebuah model mengindikasikan kemampuan model tersebut untuk memprediksi *class* target (Vercellis, 2009).

Untuk mengevaluasi model digunakan metode *confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

Confusion Matrix

Confusion matrix memberikan rincian klasifikasi, kelas yang diprediksi akan ditampilkan di bagian atas matrix dan kelas yang diobservasi ditampilkan di bagian kiri (Gorunescu, 2011). Evaluasi model *confussion matrix* menggunakan tabel seperti matrix dibawah ini:

Tabel 1. Matrik Klasifikasi untuk Model 2 *Class*

Classification		Predicted Class	
		Class = **Yes**	Class = **No**
Observed Class	Class **Yes**	(True Positive) **TP**	(False Negative) **FN**
	Class **No**	(False Positive) **FP**	(True Negative) **TN**

Sumber: Gorunescu(2011)

Akurasi dapat dihitung dengan menggunakan rumus berikut:

$$Accurasy = \frac{TP+TN}{TP+TN+FP+FN}$$

TP : Jumlah kasus positif yang diklasifikasikan sebagai positif
FP : Jumlah kasus negatif yang diklasifikasikan sebagai positif
TN : Jumlah kasus negatif yang diklasifikasikan sebagai negatif
FN : Jumlah kasus positif yang diklasidikasikan sebagai negative

Kurva ROC

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva *ROC* merupakan teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka (Gorunescu, 2011).

Kurva *ROC* adalah *tool* dua dimensi yang dipergunakan untuk menilai hasil kinerja klasifikasi yang menggunakan dua *class* sebagai keputusannya, objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP *rate* diplot pada sumbu Y dan FP *rate* diplot pada sumbu X.

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

a. 0.90-1.00 = *Excellent Classification*

b. 0.80-0.90 = *Good Classification*
c. 0.70-0.80 = *Fair Classification*
d. 0.60-0.70 = *Poor Classification*
e. 0.50-0.60 = *Failur*

The Area Under Curve (AUC) dihitung untuk mengukur perbedaan performasi metode yang digunakan. AUC dihitung menggunakan rumus (Liao & Triantaphyllou, 2007):

$$\theta^r = \frac{1}{mn}\sum_j^n = 1 \sum_i^m = 1 \psi(xi^r, xj^r$$

Dimana

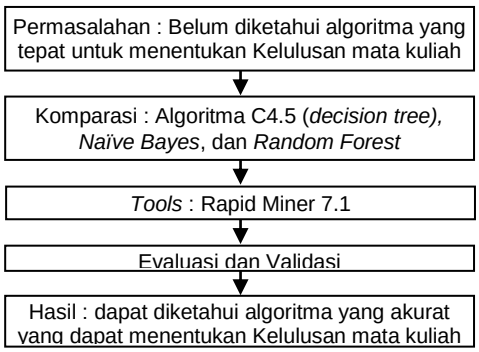
$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases}$$

X= Output Positif
Y = Output Negatif

METODE

Jenis penelitian yang digunakan dalam penelitian ini adalah model penelitian eksperimen. Penelitian ini bertujuan untuk melakukan perbandingan dan evaluasi pada algoritma klasifikasi data mining.

Penelitian eksperimen ini menekankan pada teori-teori yang sudah ada. Pada penelitian ini, jenis penelitian yang diambil adalah eksperimen komparatif, ini dilandasi oleh kerangka pemikiran pemecahan masalah seperti pada gambar 1.



Gambar 1. kerangka pemikiran pemecahan masalah

Langkah- Langkah Penelitian

Penelitian ini dilakukan dengan menjalankan beberapa langkah proses penelitian yaitu:

1. Pengumpulan data
2. Pengolahan awal data
3. Pengukuran penelitian
4. Analisa komparasi hasil

HASIL DAN PEMBAHASAN

Pengumpulan Data

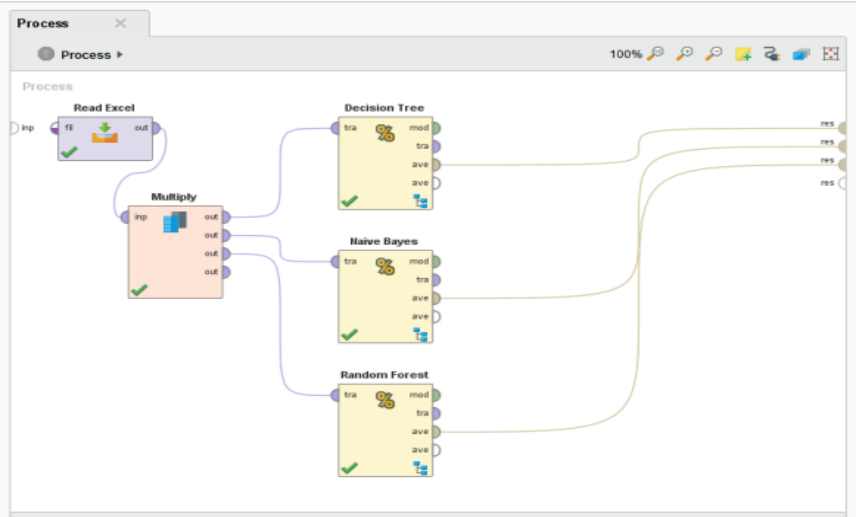
Data yang digunakan dalam penelitian ini bersumber dari Absensi mata kuliah Pemrograman 2, dimana penulis sebagai Dosen pengajar. Data merupakan hasil pemeriksaan terhadap 87 mahasiswa. Pada data ini terdiri dari 12 atribut, Seperti terlihat pada Tabel 2:

Tabel 2. Atribut dataset

No	Atribut	Tipe
1.	NPM	Integer
2.	Nama	Polynominal
3.	Jenis_Kelamin	Binominal
4.	Jenjang	Polynominal
5.	Prog_Studi	Polynominal
6.	Mata_Kuliah	Polynominal
7.	Kehadiran	Integer
8.	Tugas	Integer
9.	UTS	Integer
10.	UAS	Integer
11.	Nilai	Real
12.	Status	Binominal

Pengolahan Data Awal

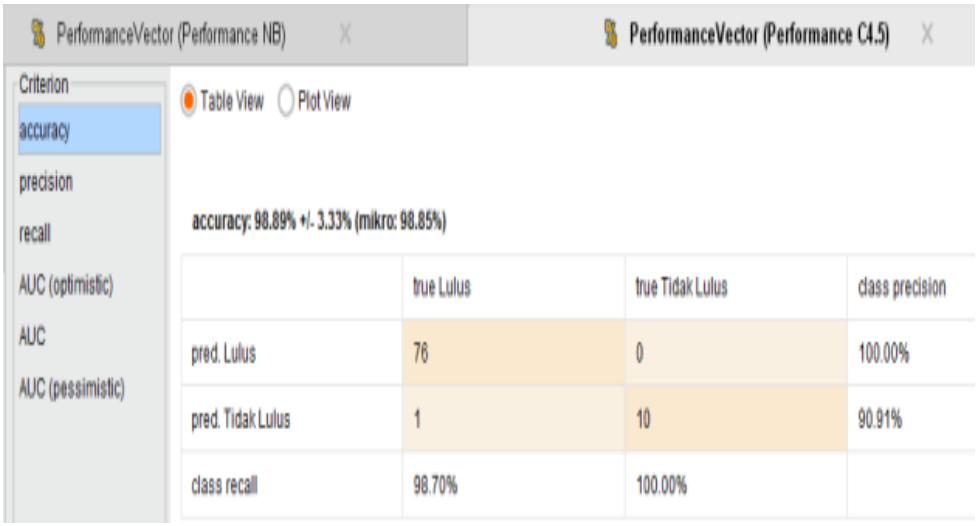
Dalam pengujian ini menggunakan *rapid miner* dengan *operator 10-fold cross-validation* untuk mendapatkan hasil *accuracy* dan AUC pada setiap algoritma yang diuji menggunakan *dataset* mahasiswa.



Gambar 2. Desain Model Komparasi Algoritma Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest

Pengukuran Penelitian

Sedangkan *Confution Matrix* guna mengukur tingkat akurasinya, yang menghasilkan nilai tertinggi dari algoritma C4.5 *(Decision Tree)*, yaitu sebesar 98.89 %, dapat kita lihat pada gambar 3 dibawah ini:

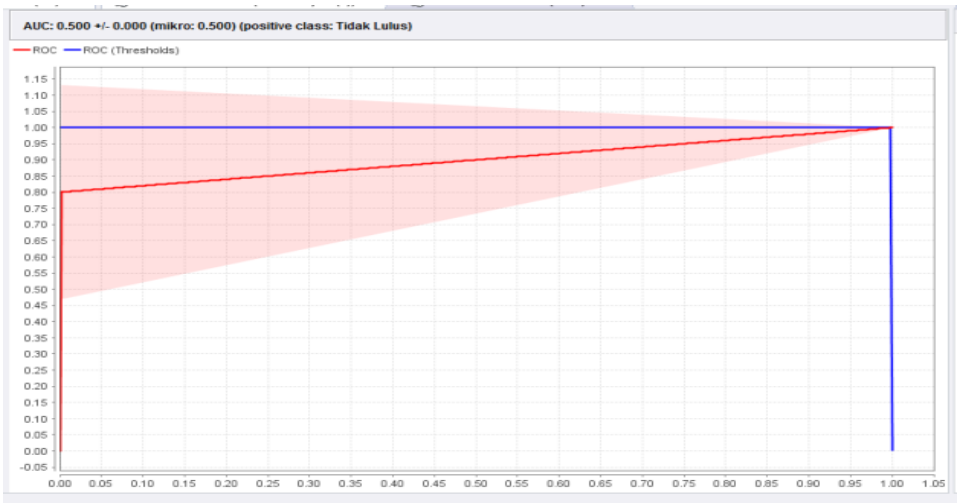


Gambar 3. Nilai Akurasi algoritma C4.5

Tabel 3. Akurasi dari semua algoritma klasifikasi

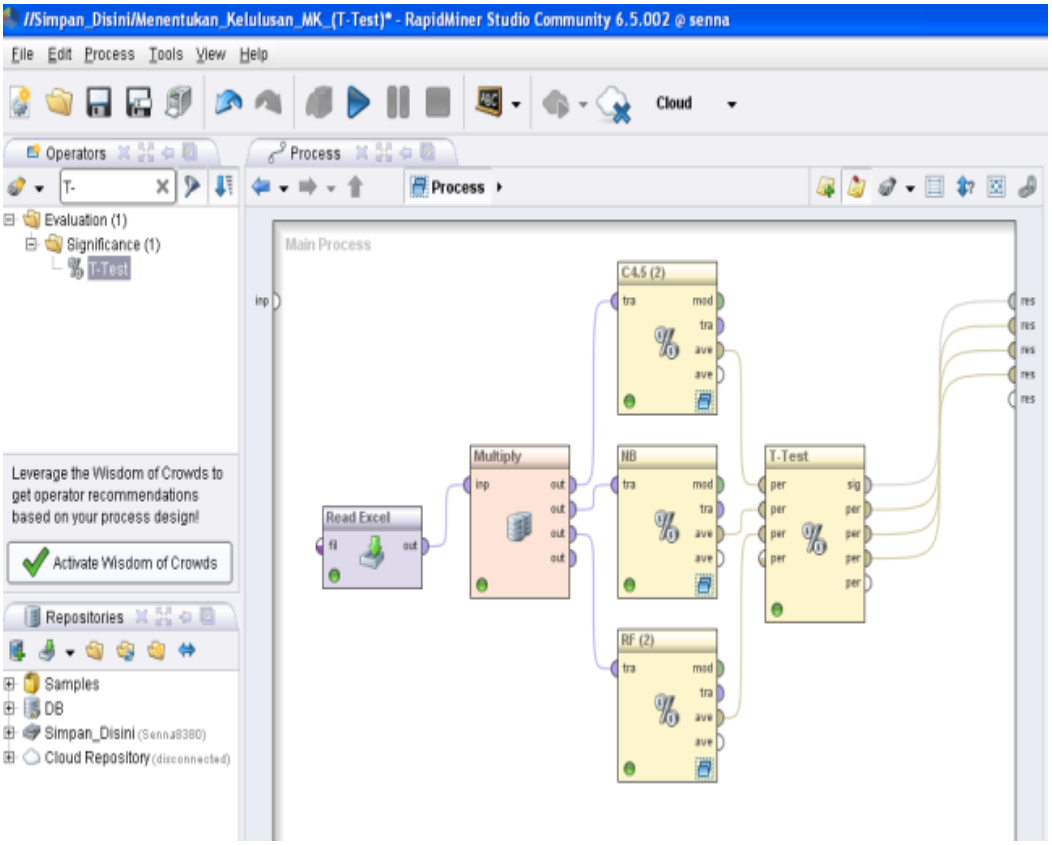
	***Confusion Matrix* (%)**
***Decision Tree* (C4.5)**	**98.89**
Naïve Bayes	**96.67**
Random Forest	**95.56**

Selanjutnya Grafik *ROC (Receiver Operating Characteristic)* dari algoritma C4.5 adalah sebgai berikut :



Gambar 4. Grafik *ROC (Receiver Operating Characteristic)*

Berdasarkan hasil evaluasi pada tabel 3, dapat dilihat bahwa algoritma yang paling baik digunakan untuk *dataset* menentukan kelulusan mata kuliah adalah *Decision Tree* (C4.5). selanjutnya dilakukan pengujian perbandingan antara masing-masing variabel yang didapat dengan menggunakan pengujian *t-test*.



Gambar 5. *T-Test*

Berikut adalah hasil dari *T-Test* yang telah dilakukan, tersaji dalam gambar 6 dibawah ini

T-Test Significance

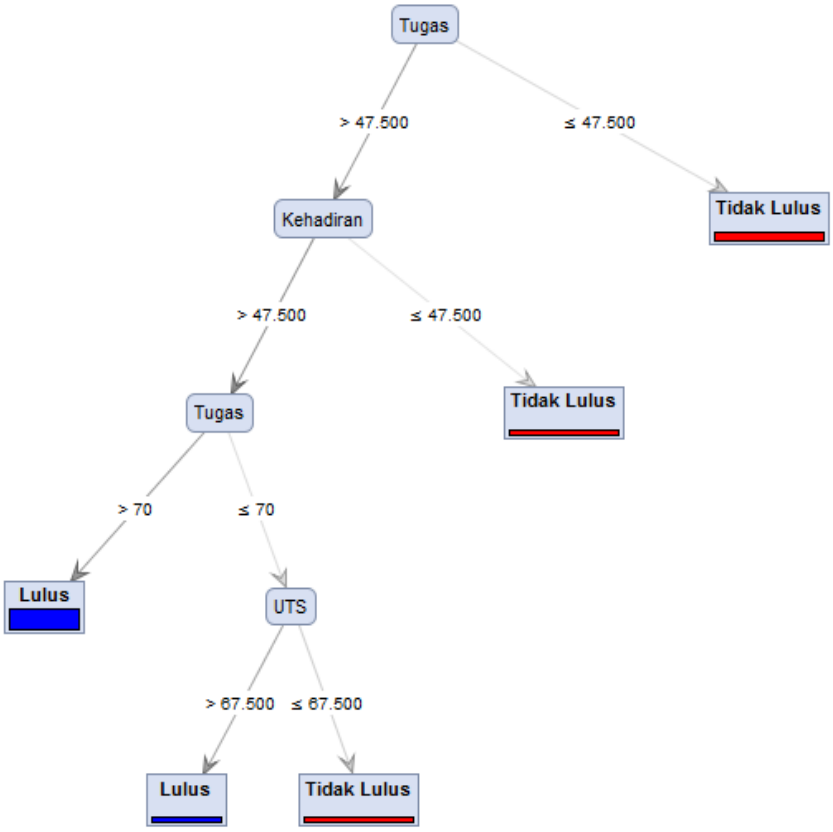
	0.978 +/- 0.044	0.956 +/- 0.054	0.954 +/- 0.056
0.978 +/- 0.044		0.491	0.454
0.956 +/- 0.054			0.956
0.954 +/- 0.056			

Probabilities for random values with the same result.
Bold values are smaller than alpha=0.050 which indicates a probably significant difference between the actual mean values!

Gambar 6. Hasil *T-Test*

Model

Melalui dataset yang disajikan pada table 1 diatas, maka dapat di buat sebuah model pohon keputusan dengan menggunakan *Rapid Miner 6.5*, berikut adalah pohon keputusan yang dihasilkan dari algoritma *decision tree* dapat dilihat pada gambar 9 :



Gambar 8. Pohon Keputusan

Dari hasil klasifikasi dengan menggunakan Algoritma C4.5, maka didapat *rule* sebagai berikut :

RuleModel

```
if Tugas > 72.500 then Lulus  (64 / 0)
if Tugas ≤ 47.500 then Tidak Lulus  (0 / 13)
if UTS > 67.500 then Lulus  (6 / 0)
else Tidak Lulus  (0 / 3)

correct: 86 out of 86 training examples.
```

Gambar 9. Rule pohon Keputusan

PENUTUP

Simpulan

Berdasarkan hasil pengujian dan analisis bahwa pengujian ini bertujuan untuk mengetahui diantara model algoritma C4.5, *Naive Bayes* dan *Random Forest* yang memiliki akurasi paling tinggi untuk menentukan kelulusan mata kuliah. Hasil perbandingan antara C4.5, *Naive Bayes* dan *Random Forest* diukur tingkat akurasinya menggunakan pengujian *Confusion Matrix* dan Kurva ROC. Berdasarkan hasil pengukuraan tingkat akurasi kedua algoritma

tersebut, diketahui bahwa nilai akurasi C4.5 (*decision tree*) adalah 98.89% dan nilai *AUC* adalah 0.500, selanjutnya nilai akurasi *Naive Bayes* 96.67% dan nilai *AUC* adalah 1.000, sedangkan nilai akurasi *Random Forest* adalah 95.56% serta nilai *UAC* adalah 1.000. Dapat disimpulkan bahwa dengan menggunakan model *C4.5 (decision tree)* lebih tinggi tingkat akurasinya, dengan peningkatan akurasi sebesar 2.22%.

DAFTAR PUSTAKA

R. Lior and M. Oded, 2015. **Data Mining With Decision Tree Theory and Applications 2nd Edition**. World Scientific Publishing Co. Pte. Ltd.

Y. Lan. et al, 2007. **Application and Comparison of Classification Techniques in Controlling Credit Risk**.

L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, 1984. **Classification and RegressionTrees, Wadsworth Statistics**, Probability Series, Belmont.

Rong Jia, Li Gang, Chen Yi-Ping P., 2009. **Accoustic Feature Selection For Automatic Emotion Recognition From Speech**, *Information Processing and Management* 45 315-328.

Bramer, M. 2007. **Principles of Data Mining** London: Springer Clark. L.A.

Kochanska, G., & Ready, R. 2000. **Mothers' personality and its interaction with child temperament as predictors of parenting behavior**. *Journal of Personality and Social Psychology*, 79, 274-285.

Urbanowicz, R. J., dan Moore, J. H. 2009. **Review Article Learning Classifier Systems: A Complete Introduction, Review, and Roadmap**. *Journal of Artificial Evolution and Applications, 2009*, 1-25.

Jang, J. S., Sun, C. T., dan Mizutani, E. 1997. **Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence**. New Jersey: Prentice Hall.