Jurnal String Vol. 3 No.1 Agustus 2018 **p-ISSN: 2527 - 9661**
e-ISSN: 2549 - 2837

IMPLEMENTASI KOMPARASI ALGORITMA KLASIFIKASI MENENTUKAN KELULUSAN MATA KULIAH ALGORITMA UNIVERSITAS BUDI LUHUR

Nahot Frastian
Program Studi Informatika, Universitas Indraprasta PGRI
nahotfrastian@gmail.com

Abstrak

Implementasi komparasi algoritma memberikan inovasi dan motivasi mahasiswa setiap semester perkuliahan. Mahasiswa Universitas Budi Luhur 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, beberapa penilaian dosen terhadap Kelulusan mata kuliah antara lain masalah kehadiran, tugas, ujian tengah semester dan ujian akhir semester. penelitian membuat algoritma yang sesuai untuk menentukan lulusan mata kuliah pada mahasiswa, peneliti ini akan menggunakan *data mining* teknik klasifikasi dengan 3 metode Algoritma klasifikasi antara lain *Algoritma C4.5 (decision tree), Naïve Bayes,* dan *Random Forest* dengan label result tidak lulus dan lulus. Hasil penelitian menguji dan melakukan dengan digunakan dataset yang sama pada ke 3 algoritma melalui perbandingan mendapat nilai *AUC* dan *Confusion Matrix,* memperoleh nilai *Area Under Curve* (*AUC*) sebesar 2.000 dari model *Naïve Bayes,* sedangkan nilai *Accuracy* atau *Confusion Matrix* yang terbesar terdapat pada algoritma C4.5 (*decision tree*) dengan nilai sebesar 98.88%. Dengan demikian, mata kuliah algoritma memberikan penilaian pada algoritma C4.5 (*decision tree*). Implementasi komparasi algoritma klasifikasi menetukan kelulusan mata kuliah algoritma di Universitas Budi Luhur.
Kata Kunci : Implementasi, komparasi algoritma, *data mining* klasifikasi, *Algoritma C4.5*

Abstract

The implementation of algorithm comparison provides innovation and motivation to students every semester of lecture. Budi Luhur University students attending all learning activities or face-to-face meetings in the classroom for each subject will have more chance to pass compared with those who rarely present. Some lecturer's evaluation to give a pass in subject include attendance, assignment, midterm examination and final examination. In this research, when creating an appropriate algorithm to determine the pass in subject of the students, this researcher will use data mining classification technique with 3 methods of Classification algorithm, namely Algorithm C4.5 (decision tree), Naïve Bayes and Random Forest with result labels of fail and pass. The results of the research tested by using the same dataset on the 3 algorithms through comparison get the value of AUC and Confusion Matrix, obtain the value of Area Under Curve (AUC) of 2,000 from the Naïve Bayes model, while the greatest Accuracy or Confusion Matrix values is in the C4.5 algorithm (decision tree) with a value of 98.88%. Thus, the algorithm subject gives an assessment of the C4.5 algorithm (decision tree). The implementation of classification algorithm comparison determines a pass in algorithm subject at Budi Luhur University.
Keywords : *Implementation, algorithm comparation, data mining classification, Algorithm C4.5*

1. PENDAHULUAN

Implementasi komparasi algoritma klasifikasi lulus kuliah tepat waktu bagi mahasiswa yang sedang mengenyam pendidikan tinggi merupakan sebuah target yang harus dicapai. Hal ini tergantung pada individu 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, karena lulus dan tidaknya mata kuliah ditentukan oleh beberapa kriteria penilaian antara lain nilai kehadiran dan nilai tugas, nilai ujian tengah semester dan nilai ujian akhir semester. Setiap mahasiswa lulus atau tidak lulus 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 menjadi *Algoritma C4.5 (decision tree), naïve bayes,* dan *Random Forest*. Begitu banyak teknologi dan metode algoritma untuk menjadi pengolah data (*data mining*), namun pada kenyataannya masih banyak dosen yang belum menerapkan teknologi dan metode algoritma, sementara data tersebut dan menjadikan metode ini penting dan bermanfaat.

Pada penelitian ini, fokus utama *data mining* yang akan dibahas adalah klasifikasi, dimana algoritma menjadi untuk klasifikasi dataset adalah *Algoritma C4.5 (decision tree), Naïve Bayes*, dan *Random Forest*. Sementara data training menjadi data penilaian mahasiswa terhadap mata kuliah Algoritma pada universitas Budi Luhur, berikut kriteria penilaian kelulusan mata kuliah antara lain nilai kehadiran dan tugas, ujian tengah semester dan nilai ujian akhir semester. Data penilaian 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 [1].

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 [2].

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.

Salah satu tujuan klasifikasi adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data [5].

C4.5 adalah algoritma yang memiliki input berupa training samples berupa data yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenaran dan *samples* yang merupakan *field - field* data yang dapat mengunakan sebagai parameter dilakukan klasifikasi data.

Algoritma dasar dari C4.5 adalah sebagai berikut:

a) Pohon yang dihasilkan berupa pohon terbalik
b) Pada tahap awal, semua contoh training adalah akar
c) Atribut adalah kategori
d) Contoh di partisi secara berulang berdasarkan atribut yang dipilih
e) Atribut tes dipilih dari data *heuristic* atau pengukuran statistik
f) Tahapan algoritma C4.5 adalah sebagai berikut:
    1. Siapkan data training
    2. Pilih atribut sebagai akar

Untuk memilih atribut akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Untuk mendapatkan nilai *Gain*, harus 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*

Decission Tree atau Pohon Keputusan adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasi. Referensi penting dalam pengerjaan aslinya adalah *Classification and Regression Tree* oleh [1].

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 membuat 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* dilakukan menguguji untuk mengecek model.

Pohon keputusan menawarkan banyak keuntungan, antara lain :

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

Naïve Bayes merupakan metode yang tidak memiliki aturan, *naïve bayes* llmu komputer yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. *naïve bayes* merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes, Simple Bayes* dan *Independence Bayes* [1].

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 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 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* menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut *subset* yang bersifat acak.

Dan dalam mengembangkan, 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 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 gunakan bahasa java dibawah lisensi GNU *Public License* dan *Rapid Miner* dapat dijalankan disistem operasi manapun. Menjadi digunakan *Rapid Miner*, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas disediakan. *Rapid Miner* dikhususkan untuk penggunaan data mining.

Validasi adalah proses mengevaluasi akurasi prediksi dari sebuah model, validasi mengacu untuk mendapatkan prediksi dengan digunakan model ada kemudian membandingkan hasil yang diperoleh dengan hasil yang diketahui [3].

Mengevaluasi akurasi dari model klasifikasi sangat penting, akurasi dari sebuah model mengindikasikan kemampuan model tersebut untuk memprediksi *class* target [6].

Untuk mengevaluasi model digunakan metode *confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*). *Confusion matrix* memberikan rincian klasifikasi, kelas yang diprediksi akan ditampilkan di bagian atas matrix dan kelas yang diobservasi ditampilkan di bagian kiri [3]. 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: [3]

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 negatif

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva *ROC* adalah

teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka [3].

Kurva *ROC* adalah *tool* dua dimensi yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua *class* keputusan, masing-masing 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 [3].

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:

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

Dimana

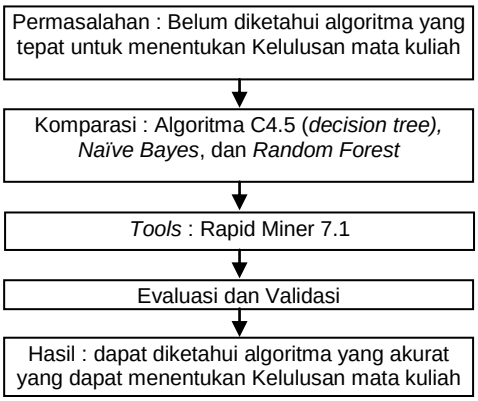
$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases}$$

X= Output Positif
Y = Output Negatif

2. METODE PENELITIAN

Metode penelitian menjelaskan 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:

a. Pengumpulan data
b. Pengolahan awal data
c. Pengukuran penelitian
d. Analisa komparasi hasil

3. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

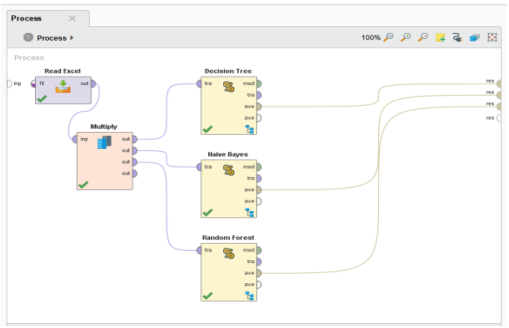
Mengumpulkan data membahas hasil peneliti bersumber dari Absensi mata kuliah Algoritma dimana penulis sebagai Dosen pengajar. Data merupakan hasil pemeriksaan terhadap 87 mahasiswa. Pada data ini terdiri dari 12 atribut.

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

B. Pengolahan Data Awal

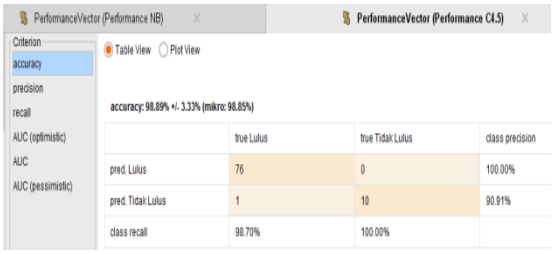
Dalam pengujian ini menggunakan *rapid miner* dengan *operator 10-fold cross-validation* untuk mendapatkan hasil *accuracy* dan AUC pada setiap algoritma yang diuji menggunakan *dataset* mahasiswa.



Gambar 3. Desain Model Komparasi Algoritma Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest

C. Pengukuran Penelitian

Sedangkan *Confution Matrix* guna mengukur tingkat akurasinya, yang menghasilkan nilai tertinggi dari algoritma C4.5 *(Decision Tree)*, yaitu sebesar 98.89 %, gambar 4 berikut:

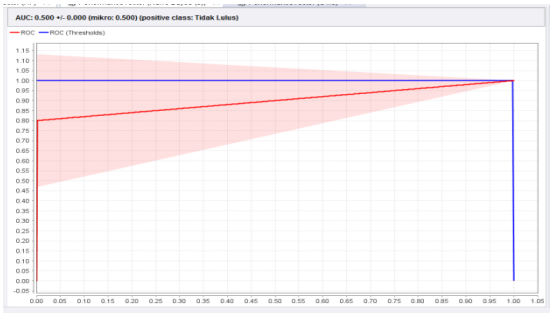


Gambar 4. Nilai Akurasi algoritma C4.5

Tabel 3. Akurasi dari Semua Algoritma Klasifikasi

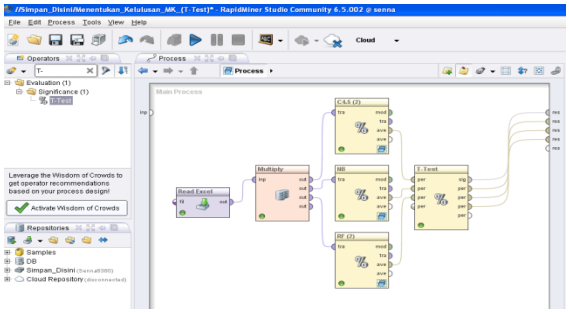
	***Confusion Matrix* (%)**
Decision Tree (C4.5)	98.89
Naïve Bayes	96.67
Random Forest	95.56

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



Gambar 5. Grafik *ROC (Receiver Operating Characteristic)*

Berdasarkan hasil evaluasi pada table 3, dapat dilihat bahwa algoritma yang paling baik digunakan untuk *dataset* ditentukan lulusan mata kuliah algoritma adalah *Decision Tree* (C4.5). selanjutnya dilakukan pengujian perbandingan antara masing-masing variabel yang didapat dengan menggunakan pengujian *t-test*. Gambar 6.



Gambar 6. *T-Test*

Hasil dari *T-Test* yang telah dilakukan, tersaji dalam gambar 7 dibawah ini :

T-Test Significance

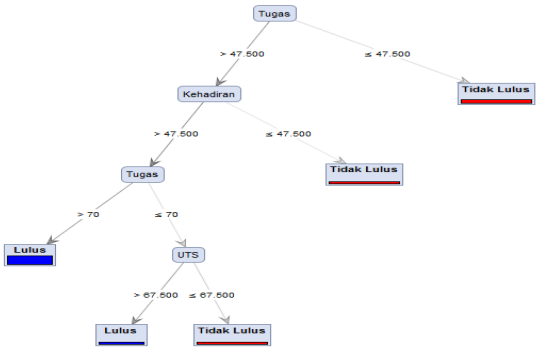
	0.978 +/- 0.044	0.956 +/- 0.054	0.954 +/- 0.056
0.978 +/- 0.044		0.491	0.454
0.956 +/- 0.054			0.956
0.954 +/- 0.056			

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 7. Hasil *T-Test*

D. 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 8 :



Gambar 8. Pohon Keputusan

Dari hasil klasifikasi dengan menggunakan Algoritma C4.5, maka didapat rule sebagai berikut Gambar 9 :

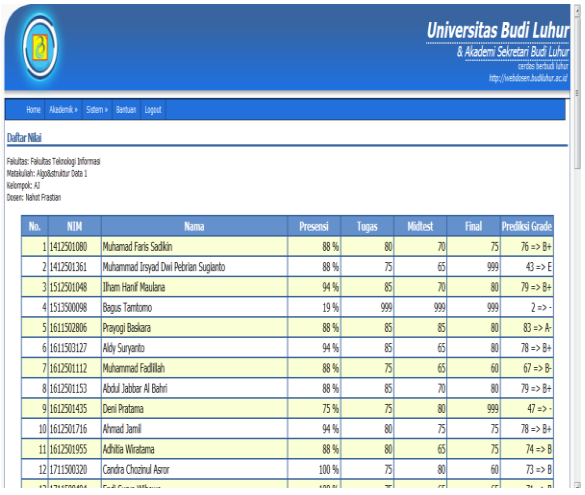
RuleModel

```
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

Berdasarkan hasil evaluasi pada rule pohon keputusan, dapat dilihat bahwa algoritma yang paling baik digunakan untuk *data* ditentukan lulusan mata kuliah Algoritma di Universitas Budi Luhur. selanjutnya dilakukan pengujian perbandingan antara masing-masing variabel yang didapat dengan menggunakan pengujian. Gambar 10.



No.	NIM	Nama	Presensi	Tugas	Midtest	Final	Prediksi Grade
1	1412501080	Muhamad Faris Sadikin	88 %	80	70	75	76 => B+
2	1412501361	Muhammad Irsyad Dwi Pebrian Suglanto	88 %	75	65	999	43 => E
3	1512501048	Ilham Hanif Maulana	94 %	85	70	80	79 => B+
4	1513500098	Bagus Tamtomo	19 %	999	999	999	2 => -
5	1611502806	Prayogi Baskara	88 %	85	85	80	83 => A-
6	1611503127	Aldy Suryanto	94 %	85	65	80	78 => B+
7	1612501112	Muhammad Fadillah	88 %	75	65	60	67 => B-
8	1612501153	Abdul Jabbar Al Bahri	88 %	85	70	80	79 => B+
9	1612501435	Deni Pratama	75 %	75	80	999	47 => -
10	1612501716	Ahmad Jamil	94 %	80	75	75	78 => B+
11	1612501955	Achitia Wiratama	88 %	80	65	75	74 => B
12	1711500320	Candra Chozinul Asror	100 %	75	80	60	73 => B

Gambar 10. Penilaian mata kuliah Algoritma Universitas Budi Luhur

4. SIMPULAN

Bahwasanya hasil pengujian dan analisis menjadi tujuan untuk mengetahui antara model algoritma C4.5, *Naive Bayes* dan *Random Forest* yang memiliki akurasi paling tinggi untuk ditentukan lulusan mata kuliah algoritma. 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* 97.67% dan nilai *AUC* adalah 2.000, sedangkan nilai akurasi *Random Forest* adalah 96.56% serta nilai *UAC* adalah 2.000. Dapat disimpulkan bahwa dengan menggunakan model *C4.5 (decision tree)* lebih tinggi tingkat akurasinya, menjadi meningkat akurasi sebesar 2.22%.

DAFTAR PUSTAKA

[1] 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.

[2] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

[3] Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques.* Berlin: Springer.

[4] 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.

[5] Khan, I. A., dan Choi, J. T. (2014). An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction. *International Journal of Software Engineering and Its Applications, 8*(12), 31-42.

[6] Vercellis, C. (2009). *Business Intelligence.* United John Wiley and Sons.