KOMPARASI ALGORITMA C4.5, NAIVE BAYES, DAN RANDOM FOREST UNTUK KLASIFIKASI DATA KELULUSAN MAHASISWA JAKARTA

Oleh:
Ibnu Alfarobi, Taransa Agasya Tutupoly, Ade Suryanto
AMIK BSI Tangerang
Email: robi.alfa.ibnu@gmail.ac.id

ABSTRAK

Mengetahui tingkat kelulusan mahasiswa dalam institusi pendidikan sangatlah penting. Selain untuk menjaga kredibilitas institusi tersebut, juga berperan dalam menjaga rasio antara mahasiswa dengan dosen agar tetap dalam takaran yang tepat.Salah satu disiplin ilmu pengetahuan yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar adalah Data Mining. Penelitian ini dilakukan dengan membagi data testing dan data training dengan perbandingan 10 : 90, 20 : 80, dan 30 : 70. Tujuan penelitian ini untuk mengkomparasikan algoritma C4.5, Naive Bayes, dan Random Forest dalam penentuan klasifikasi data kelulusan mahasiswa. Hasil penelitian menunjukkan bahwa secara keseluruhan algoritma C4.5 mempunyai akurasi paling besar jika dibandingkan dengan algoritma lainnya dengan tingkat akurasi sebesar 85.34% pada eksperimen pertama dan 89.06% pada ekperimen ketiga. Sedangkan pengukuran dengan menggunakan ROC curve, algoritma Naive Bayes menjadi algoritma yang mempunyai tingkat akurasi teringgi dibandingkan dengan algoritma C4.5 dan Random Forest dengan nilai AUC sebesar 0.925.

Kata kunci: Klasifikasi, Kelulusan Mahasiswa, C4.5, Naive Bayes, Random Forest.

PENDAHULUAN

Latar Belakang

Mengetahui tingkat kelulusan mahasiswa dalam suatu institusi pendidikan sangatlah penting. Selain untuk tetap menjaga kredibilitas institusi tersebut, tingkat kelulusan juga berperan dalam menjaga rasio antara mahasiswa dengan dosen agar tetap dalam takaran yang tepat. Untuk itu, informasi yang cepat, tepat, dan akurat tentang klasifikasi tingkat kelulusan mahasiswa akan sangat dibutuhkan supaya pihak institusi dapat membuat strategi ataupun solusi yang tepat agar dapat menjaga bahkan meningkatkan *trend* positif terkait tingkat kelulusan mahasiswa.

Saat ini sebuah perguruan tinggi atau Universitas dituntut untuk selalu memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang dimilikinya. Teknologi yang berkembang sampai saat ini pun membuat sebuah sistem informasi berperan semakin penting dalam menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis.

Salah satu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar adalah *Data Mining*.

Data mining adalah proses melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data(Witten et al., 2011). *Huge of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data* (Han et al., 2011). *Data mining*sering dianggap sebagai bagian dari *Knowledge Discovery in Database* (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. Selain itu *data mining* juga dikenal dengan nama *knowledge extraction, pattern analysis, information harvesting,* dan *Business intelligence.*

Ada 5 peranan utama *data mining,* yaitu: Estimasi, Prediksi, Klasifikasi, Klastering, dan Asosiasi. Algoritma data mining yang sering digunakan dalam klasifikasi diantaranya adalah Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, dan lain-lain. Namun, pada Tesis ini penulis hanya akan menggunakan algoritma C4.5, Naive Bayes, dan Random Forest untuk mengolah, mengklasifikasikan, serta *memining knowledge* dari dataset kelulusan mahasiswa.

Pemilihan penggunaan algoritma C4.5, Naive Bayes, dan Random Forest pada penelitian ini didasarkan pada beberapa alasan, yaitu: Selain ketiga algoritma tersebut sama-sama mudah diimplementasikan dan sama-sama dapat memberikan hasil yang baik dalam kasus klasifikasi, ketiga algoritma tersebut juga mempunyai beberapa keunggulan masing-masing. C4.5 merupakan algoritma klasifikasi pohon keputusan yang efisien dalam menangani atribut bertipe diskret dan numerik (Han et al., 2012). Algoritma Naive Bayes, (Han et al., 2012) menjelaskan bahwa algoritma ini hanya membutuhkan satu kali scan data training.

Dalam *data mining,* penelitian mengenai klasifikasi kelulusan mahasiswa sudah pernah dilakukan oleh peneliti lain. Sebagian besar penelitian tersebut difokuskan pada identifikasi variabel prediktor. Ada banyak penelitian dalam literatur terdahulu yang menjelaskan faktor-faktor apa saja yang dapat mensukseskan proses pengklasifikasian kelulusan mahasiswa. Faktor-faktor tersebut secara umum dibagi menjadi dua, yaitu faktor pra penerimaan mahasiswa dan pasca penerimaan mahasiswa. Prestasi SMA, peringkat SMA adalah prediktor kelulusan yang lebih baik dari pada nilai tes masuk perguruan tinggi (Niu & Tienda, 2009), peneliti lain menemukan korelasi antara kualitas SMA dengan keberhasilan siswa di perguruan tinggi mempengaruhi kelulusannya (Black et al., 2015).

Database yang ada di dunia pada saat ini sangat rentan terhadap *noisy data,* data yang hilang atau tidak lengkap, dan data yang tidak konsisten karena biasanya ukuran dari database tersebut sangat besar serta sumber dari data-data tersebut biasanya lebih dari satu (heterogen). Untuk itu, menyiapkan data yang baik, memadai dan representatif merupakan langkah awal yang tidak dapat diabaikan begitu saja. Kehandalan informasi yang akan di *mining* dari sebuah database yang ada bergantung pada kualitas data yang nantinya akan diproses. Ada beberapa teknik *data preprocessing* yang dapat digunakan untuk menghasilkan data yang berkualitas. *Data cleaning* dapat diterapkan untuk menghilangkan *noise* dan data yang tidak konsisten. *Data integration* dapat digunakan untuk menggabungkan data-data dari banyak sumber menjadi satu data yang saling berhubungan dalam satu *data store. Data reduction* dapat mengurangi ukuran data. *Data*

transformation untuk meningkatkan akurasi dan efisiensi algoritma *mining* yang melibatkan pengukuran jarak (Han, 2012).

Rumusan Maslah

Penggunaan algoritma C4.5, *Naive Bayes,* dan *Random Forest*sudah pernah digunakan untuk mengolah dan *memining knowledge* dari dataset kelulusan mahasiswa. Ketiga algoritma *data mining* tersebut pun masing-masing mempunyai kelebihan dan kekurangan. Namun dari ketiganya belum dapat dipastikan model mana yang lebih akurat dan cepat dalam melakukan klasifikasi. Hal ini dikarenakan dataset yang digunakan oleh peneliti sebelumnya tidak sama, perlakuan yang diberikan kepada data tersebut juga berbeda. Semakin kompleks data, *noise* pada data, serta data yang tidak konsisten tentunya akan berdampak pada kinerja algoritma pengklasifikasinya. Untuk mendapatkan model algoritma yang paling baik, maka penulis membandingkan tiga algoritma di atas dengan menggunakan dataset yang sama serta perlakuan yang sama pada dataset tersebut.

Berdasarkan identifikasi masalah di atas, maka pada penelitian ini berusaha menjawab pertanyaan model mana yang lebih akurat dan cepat antara algoritma C4.5, *Naive Bayes,* dan *Random Forest* dalam klasifikasi kelulusan mahasiswa.

Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membandingkan algoritma yang paling akurat dalam penentuan klasifikasi kelulusan mahasiswa. Algoritma-algoritma yang digunakan adalah C4.5, *Naive Bayes,* dan *Random Forest.*

KAJIAN PUSTAKA

Data Mining

Salah satu disiplin ilmu yang dapatdigunakan untuk menemukan pola atau *memining knowledge* dari suatu *big data* yang ada adalah *Data Mining. Data mining* seringdianggap sebagai bagian dari *Knowledge Discovery in Database* (KDD) yaitu sebuah proses mencaripengetahuan yang bermanfaat dari data atau ekstrasi pola secara otomatis mewakili pengetahuan yang disimpan atau ditangkapsecara tersembunyi didalam sebuah database besar, gudang data, web, repositoriinformasi lainnya, atau data stream (Han et al., 2012).

Pada dasarnya, *data mining* dapat dilihat sebagai ilmu yang mengeksplorasi dataset dalam jumlahbesar untuk penggalian informasi yang tersirat, yang sebelumnya tidak diketahuidan berpotensi menghasilkan informasi yang berguna (Gorunescu, 2011). *Data mining* adalah prosesterorganisir untuk mengidentifikasi pola yang valid, baru, berguna, dan dapatdimengerti dari sebuah dataset yang besar dan kompleks (Maimon & Rokach,2010).

Klasifikasi

Klasifikasi adalah proses menempatkan obyek atau konsep tertentu kedalam satu set kategori, berdasarkan sifat obyek ataukonsep yang bersangkutan (Gorunescu, 2011). Dalam klasifikasi terdapat dua pekerjaan utama yang dilakukan: pertama, pembangunan model sebagai *prototype* untuk disimpan sebagai memori. Kedua, penggunaan model tersebut untuk melakukan

pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut berada.Proses klasifikasi didasarkan pada komponen (Gorunescu, 2011):

1. Kelas (*Class*)
   Variabel dependen dari model yang merupakan kategori variabel yang mewakili label-label yang diletakkan pada obyek setelah pengklasifikasian. Contoh: kelas bintang, kelas gempa bumi
2. Prediktor (*Predictor*)
   Variabel independen dari model yang diwakili oleh karakteristik atau atribut dari data yang diklasifikasikan berdasarkan klasifikasi yang dibuat. Contoh: tekanan darah, status perkawinan, musim
3. Dataset Pelatihan (*Training Dataset*)
   Merupakan dataset yang berisi dua komponen nilai yang digunakan untuk pelatihan mengenali model yang sesuai dengan kelasnya, berdasarkan prediktor yang ada. Contoh: database penelitian gempa, database badai, database pelanggan supermarket
4. Database Pengujian (*Testing Database*)
   Merupakan dataset baru yang akan diklasifikasikan oleh model yang dibangun sehingga dapat dievaluasi hasil akurasi klasifikasi tersebut

Algoritma C4.5

Salah satu motode klasifikasi yang melibatkan konstruksi pohon keputusan, koleksi node keputusan, terhubung oleh cabang-cabang, memperpanjang ke bawah dari simpul akar sampai berakhir di node daun. Dimulai dari node root, yang oleh konvensi ditempatkan dibagian atas dari diagram pohon keputusan, atribut diuji pada node keputusan, dengan setiap hasil yang mungkin menghasilkan cabang. Setiap cabang kemudian mengarah ke node lain baik keputusan atau ke node daun untuk mengakhiri.

Algoritma C4.5 dan pohon keputusan (*decision tree*) merupakan dua mode yang tidak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5. Decision Tree merupakan algoritma pengklasifikasian yang sering digunakan dan mempunyai struktur yang sederhana dan mudah untuk diinterpretasikan (Mantas & Abellan, 2014).

Naive Bayes

Naive Bayes merupakan salah satu metode*machine learning* yang menggunakan perhitunganprobabilitas. Algoritma ini memanfaatkan metodeprobabilitas dan statistik yang dikemukakan olehilmuwan Inggris bernama Thomas Bayes, yaitumemprediksi probabilitas di masa depan berdasarkanpengalaman di masa sebelumnya. Algoritma pengklasifikasi Naive Bayes adalah pengklasifikasi yangberdasarkan probabilitas bersyarat pada teorema Bayes (Aggarwal, 2015).

Random Forest

Random Forest merupakan pengembangandari Decision Tree, dimana setiap Decision Tree telahdilakukan training menggunakan sampel individu dansetiap

atribut dipecah pada tree yang dipilih antaraatribut subset yang bersifat acak. Dan pada prosesklasifikasi, individunya didasarkan pada vote darisuara terbanyak pada kumpulan populasi tree.

Kelulusan Mahasiswa

Mahasiswa dapat didefinisikansebagai individu yang sedang menuntut ilmu ditingkat perguruan tinggi,baik negeri maupun swasta atau lembaga lain yang setingkat denganperguruan tinggi. Pemantauan mahasiswa yang masuk, peningkatan kemampuan mahasiswa, prestasi yang dicapai mahasiswa, rasio kelulusan seharusnya menjadi perhatian yang sangat serius bagi setiap perguruan tinggi karena merupakan satuan pendidikan yang menjadi terminal akhir bagi setiap orang yang ingin menuju kejenjang pendidikan yang lebih tinggi.

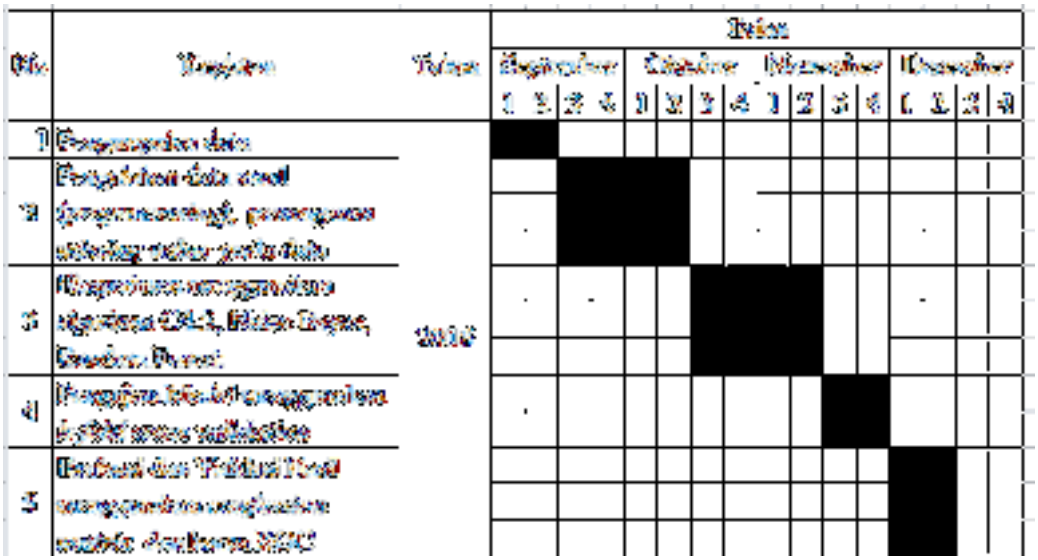
METODE PENELITIAN

Desain Penelitian

Penelitian adalah usaha mencari melalui proses yang metodis untuk menambahkan pengetahuan itu sendiri dan dengan yang lainnya, oleh penemuan fakta dan wawasan tidak biasa (Dawson, 2009). Untuk dapat menemukan fakta atau pengetahuan dari data, dibutuhkan suatu usaha ekstraksi yang disebut dengan *data mining.*

Menurut Dawson dalam Setiyorini et al, terdapat beberapa metode penelitian yang dapat dipakai untuk mengatasi masalah penelitian yaitu *action research, experiment, case study* dan *survey* (Setiyorini et al., 2014). Dalam penelitian ini, metode penelitian yang digunakan adalah metode penelitian eksperimen dengan tahapan: pengumpulan data, pengolahan data awal, eksperimen, pengujian model, evaluasi dan validasi hasil.

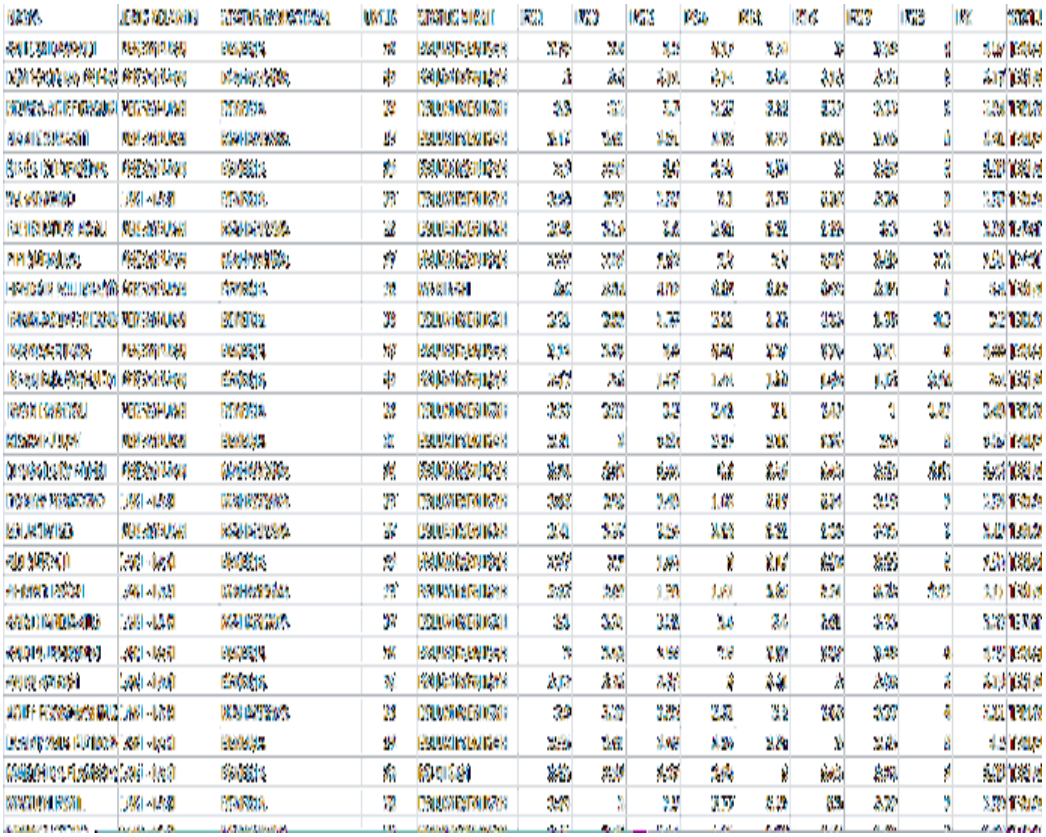
Tabel 1: Jadwal Implementasi



Pengumpulan Data

Data yang peneliti ambil merupakan data kelulusan mahasiswa yang mempunyai 379 *record* dan terdiri dari 15 atribut yaitu nama, status mahasiswa, umur, status nikah, IPS 1, IPS 2, IPS 3, IPS 4, IPS 5, IPS 6, IPS 7, IPS 8, IPK, dan status kelulusan. Sebagai contoh data kelulusan mahasiswa yang belum diolah dapat dilihat pada tabel 1 dibawah ini.

Tabel 2: Data Kelulusan Mahasiswa



Sumber :http://romisatriawahono.net/lecture/dm/dataset/

Pengolahan Data Awal

Tahapan selanjutnya adalah pengolahan data awal, setelah data terkumpul maka diperlukan *preprocessing* data terlebih dulu. Hal ini bertujuan untuk membersihkan dataset yang sudah ada dari data-data yang tidak perlu. Dataset yang digunakan dalam penelitian ini, masih ditemukan mempunyai *missing value* yang harus diperlakukan secara khusus. Adapun penanganan *missing value* menurut (Han et al., 2012) adalah:

1. Mengabaikan *tuple* yang berisi *missing value*
2. Mengganti *missing value* secara manual
3. Mengganti *missing value* dengan konstanta global (misal "*unknown"* atau "∞")
4. Mengganti *missing value* dengan nilai *mean* atau *median* dari atribut
5. Mengganti *missing value* dengan nilai *mean* atau *median* dari semua sampel
6. Mengganti *missing value* dengan nilai kemungkinan terbanyak dari dataset

Pada penelitian ini, perlakuan khusus yang diberikan untuk menangani *missing value* adalah dengan memberikan nilai rata-rata dari atribut. Teknik ini dapat diterapkan untuk atribut yang mempunyai nilai numerik.

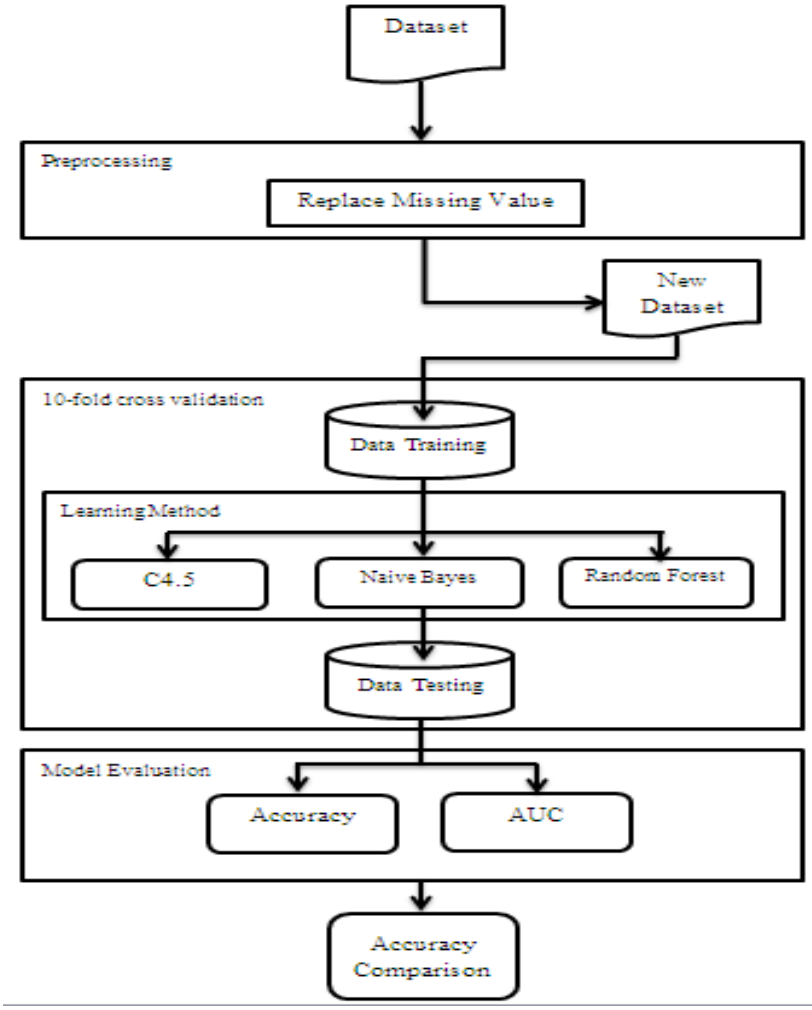
Pengujian Model

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tiga metode klasifikasi data mining. Algoritma yang akan digunakan adalah C4.5, Naive Bayes, dan Random Forest. Setelah diolah dan menghasilkan model, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menggunakan *k-fold cross validation* dengan perbandingan antara *data testing* dan *data training* 10 : 90, 20 : 80, 30 : 70 dan mengulang pengujian tersebut beberapa kali.

Evaluasi dan Validasi Hasil

Tahap selanjutnya adalah melakukan evaluasi dan validasi hasil pengujian model tersebut dengan menggunakan *confussion matrix* dan kurva ROC. *Confussion matrix* adalah alat (*tools*) visualisasi yang biasa digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda (Han et al., 2012). Sedangkan kurva ROC menurut (Attenberg & Ertekin, 2013) adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif. Untuk mengolah dataset yang ada, akan digunakan metode yang diusulkan seperti yang terlihat pada gambar 2 berikut:

Gambar 1: Model yang diusulkan



HASIL PENELITIAN DAN PEMBAHASAN

Metode klasifikasi bisa dievaluasi berdasarkan kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas dan interpretabilitas (Vecellis, 2009). Pada penelitian ini, eksperimen yang dilakukan bertujuan untuk mengetahui tingkat akurasi yang terbaik diantara algoritma C4.5, Naive Bayes, dan Random Forest dengan membandingkan ketiga algoritma tersebut. Setelah diolah dan menghasilkan model, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian dengan menggunakan *k-fold cross validation* dengan perbandingan antara *data testing* dan *data training* nya yaitu sebagai berikut:10 : 90, 20 : 80, 30 : 70.

Hasil Penelitian

Hasil Penelitiandan Pengujian Model C4.5

Model *confussion matrix* akan membentuk matrix yang terdiri dari *true positive* atau tupel positif dan *true negative* atau tupel negatif. Dari sebanyak 379 data kelulusan mahasiswa yang telah diolah menggunakan algoritma C4.5 di rapidminer dengan perbandingan data testing dan data training 10% : 90%, terdapat sebanyak 128 data yang di prediksi terlambat dan kenyataannya terlambat, 163 data diprediksi tepat dan kenyataannya tepat, 32 data diprediksi terlambat tetapi kenyataannya tepat, 18 data diprediksi tepat namun kenyataannya terlambat seperti pada gambar 2. Untuk perbandingan data testing dan data training 20% : 80%, terdapat sebanyak 110 data yang di prediksi terlambat dan kenyataannya terlambat, 148 data diprediksi tepat dan kenyataannya tepat, 29 data diprediksi terlambat tetapi kenyataannya tepat, 16 data diprediksi tepat namun kenyataannya terlambat. Sedangkan Untuk perbandingan data testing dan data training 30% : 70%, terdapat sebanyak 92 data yang di prediksi terlambat dan kenyataannya terlambat, 144 data diprediksi tepat dan kenyataannya tepat, 13 data diprediksi terlambat tetapi kenyataannya tepat, 16 data diprediksi tepat namun kenyataannya terlambat.

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC)yang didapat dengan menggunakan algoritma C4.5 serta perbandingan data testing dan data trainingnya adalah 10% : 90% menghasilkan nilai AUC sebesar = 0.856 seperti pada gambar 3, perbandingan data testing dan data trainingnya adalah 20% : 80% menghasilkan nilai AUC sebesar = 0.834, dan untuk perbandingan data testing dan data trainingnya adalah 30% : 70% menghasilkan nilai AUC sebesar = 0.869.

Gambar 2: *Confussion Matrix* Algoritma C4.5
(Untuk pembagian data testing 10% berbanding data training 90%)

Sumber : Data Hasil Penelitain diolah (2016)

Nilai *Accuracy* adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

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

$$= \frac{163 + 128}{163 + 128 + 32 + 18}$$

$$= \frac{291}{341}$$

$$= 0.85337$$

$$= 85.34\%$$

Keterangan:
TP = *True Positive*
TN = *True Negative*
FP = *False Positive*
FN = *False Negative*

Gambar 3 : Grafik Area Under Curve (AUC) Algoritma C4.5
(Untuk pembagian data testing 10% berbanding data training 90%)

Sumber : Data Hasil Penelitain diolah (2016)

Kriteria penilaian (Gorunescu, 2011) :

1. 0.90 - 1.00 = excellent classification
2. 0.80 - 0.90 = good classification
3. 0.70 - 0.80 = fair classification
4. 0.60 - 0.70 = poor classification
5. 0.50 - 0.60 = failure

Hasil Penelitian dan Pengujian Model Naive Bayes

Nilai akurasi yang diperoleh dengan menggunakan algoritma naive bayes dengan perbandingan data testing 10% : data trainingnya 90% adalah; *accuracy* = 85.34% dan *Area Under Curve* (AUC) adalah 0.823. Dari keselurahan 379 dataset yang diolah, sebanyak 117 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 174 data diprediksi tepat dan pada kenyataannya memang tepat, 21 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 29 data diprediksi tepat tetapi kenyataannya terlambat. Perbandingan data testing 20% : data trainingnya 80% adalah; *accuracy* = 83.83% seperti pada gambar 4.11 dan *Area Under Curve* (AUC) adalah 0.907. Dari keselurahan 379 dataset yang diolah, sebanyak 107 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 147 data diprediksi tepat dan pada kenyataannya memang tepat, 30 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 19 data diprediksi tepat tetapi kenyataannya terlambat. Perbandingan data testing 30% : data trainingnya 70% adalah; *accuracy* = 86.79% seperti pada gambar 4.12 dan *Area Under Curve* (AUC) adalah 0.925 seperti pada gambar 4.15. Dari

keselurahan 379 dataset yang diolah, sebanyak 96 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 134 data diprediksi tepat dan pada kenyataannya memang tepat, 23 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 12 data diprediksi tepat tetapi kenyataannya terlambat.

Hasil Penelitian dan Pengujian Model Random Forest

Nilai akurasi yang diperoleh dengan menggunakan algoritma random forest serta perbandingan data testing 10% : data training 90% adalah; *accuracy* = 73.61% seperti pada gambar 4.16 dan *Area Under Curve* (AUC) adalah 0.823. Dari keselurahan 379 dataset yang diolah, sebanyak 72 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 179 data diprediksi tepat dan pada kenyataannya memang tepat, 16 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 74 data diprediksi tepat tetapi kenyataannya terlambat. Perbandingan data testing 20% : data training 80% adalah; *accuracy* = 85.81% seperti pada gambar 4.17 dan *Area Under Curve* (AUC) adalah 0.886 seperti pada gambar 4.20. Dari keselurahan 379 dataset yang diolah, sebanyak 100 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 160 data diprediksi tepat dan pada kenyataannya memang tepat, 17 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 26 data diprediksi tepat tetapi kenyataannya terlambat. Perbandingan data testing 30% : data training 70% adalah; *accuracy* = 76.23% seperti pada gambar 4.18 dan *Area Under Curve* (AUC) adalah 0.842. Dari keselurahan 379 dataset yang diolah, sebanyak 74 jumlah data yang diprediksi terlambat dan pada kenyataannya memang terlambat, 128 data diprediksi tepat dan pada kenyataannya memang tepat, 29 data yang diprediksi terlambat tetapi kenyataannya tepat, dan 34 data diprediksi tepat tetapi kenyataannya terlambat.

Pembahasan

Untuk rata-rata keseluruhan percobaan dapat dilihat pada tabel 2 berikut ini.

Tabel 3 Rata-rata Hasil Komparasi Algoritma Berdasarkan Data Testing dan Data Training

Algoritma	Data Testing	Data Training	Accuracy	AUC
C4.5	10	90	85.34%	0.846
	20	80	85.15%	0.834
	30	70	89.06%	0.869
	Rata - Rata		**86.52%**	**0.850**
Naive Bayes	10	90	85.34%	0.823
	20	80	83.83%	0.907
	30	70	86.79%	0.925
	Rata - Rata		**85.32**	**0.885**
Random Forest	10	90	73.61%	0.823
	20	80	85.81%	0.886
	30	70	76.23%	0.842
	Rata - Rata		**78.55**	**0.850**

Pada tabel 3, dapat kita lihat bahwa rata-rata akurasi dari algoritma C4.5 adalah 86.52 %, ini adalah rata-rata akurasi yang paling tinggi jika dibandingkan dengan Naive Bayes dan Random Forest. Hal tersebut dikarenakan algoritma C4.5 memang mempunyai struktur yang sederhana dan mudah untuk diinterpretasikan (Mantas & Abellan, 2014) serta mudah untuk dimengerti meskipun oleh pengguna yang belum ahli sekalipun dan lebih efisien dalam menginduksi data (C. Sammut, 2011).

KESIMPULAN DAN SARAN

Kesimpulan

Dari hasil komparasi algoritma C4.5, Naive Bayes, dan Random Forest, dari percobaaan dengan pembagian *data testing* : *data training* 10 : 90, 20 : 80, 30 : 70. Jika dibandingkan dengan nilai akurasi algoritma naive bayes dan algoritma random forest, nilai akurasi dengan menggunakan algoritma klasifikasi C4.5 adalah yang terbesar pada percobaan data testing 10% : data training 90% dan percobaan data testing 30% : data training 70%. Sedangkan evaluasi menggunakan ROC *curve* yaitu berdasarkan nilai AUC, algoritma naive bayes menjadi yang tertinggi pada percobaan data testing 20% : data training 80% dan data testing 30% : data training 70% dengan nilai mendekati 1.000 yaitu 0.907 dan 0.925. Dari hasil keseluruhan pengujian model dapat disimpulkan bahwa kinerja C4.5 dan Naive

Bayes hampir sama bagusnya, baik itu dilihat dari tingkat akurasi maupun AUC nya.

Saran

Untuk keperluan penelitian lebih lanjut mengenai komparasi metode klasifikasi data mining, dapat dilakukan pengembangan untuk dapat menghasilkan model yang lebih baik lagi, diantaranya:

1. Untuk mendapatkan nilai akurasi yang lebih baik lagi, dapat digunakan operator optimasi seperti *Particle Swarm Optimization* (PSO), *Ant Colony Optimization* (ANT), *Genetik Algorithm* (GA), *Chi Square*, dan lain sebagainya.
2. Eksperimen penelitian dapat menggunakan jumlah data yang lebih banyak lagi dan menguji coba dengan dataset kelulusan mahasiswa yang lain sehingga model yang sudah didapat akan lebih teruji lagi.
3. Menggunakan algoritma pengklasifikasi lain yang mungkin diluar *supervised learning* agar dapat dilakukan penelitian yang berbeda dari umumnya yang sudah ada

DAFTAR PUSTAKA

Anggarwal, Charu C. (2015). *Data Mining: The Textbook.* New York: Springer.

Blaxter, L., Hughes, C., & Tight, M. (2010). *How to Research* (4th ed). Maidenhead: Open University Press.

Dawson, C. W. (2009). *Projects in Computing and Information Systems a student's guide*. Harlow, UK: Addison-Wesley.

Gorunescu, Florin (2011). *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer.

Han, J., &Kamber.,& Pei, J. (2012). Data Mining Consepts and Techniques. San Fransisco: Morgan Kauffman.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery. Handbook.* London: Springer.

Mantas, C. J., & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, *41*(10), 4625–4637. doi:10.1016/j.eswa.2014.01.017.

Sammut, Claude. (2011). *Encyclopedia of Machine Learning*. Boston, MA: Springer.

Setiyorini, T., Pascasarjana, P., Ilmu, M., Tinggi, S., Informatika, M., Komputer,D. a N., & Mandiri, N. (2014a). Penerapan Metode Bagging UntukMengurangi Data Noise Pada Neural Network Untuk Estimasi Kuat TekanBeton Penerapan Metode Bagging Untuk Mengurangi Data Noise Pada NeuralNetwork Untuk, 1(1), 36–41.

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

W. C.-M. Liaw, Yi-Ching, Leou Maw-Lin, "*Fast exact k nearest neighbors search using anorthogonal search tree," Pattern Recognit*., vol. 43, no. 6, pp.2351–2358, Feb. 2010.

Witten, I. H., Frank, E., & Hall, M. A. (2011).*Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.