Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree

Achmad Bisri
Fakultas Teknik, Universitas Pamulang
Email: achmadbizri@gmail.com

Romi Satria Wahono
Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Email: romi@brainmatics.com

Abstract: Universitas Pamulang salah satu perguruan tinggi yang memiliki jumlah mahasiswa yang besar, namun dalam data histori terdapat masalah dengan jumlah kelulusan yang tepat waktu dan terlambat (tidak tepat waktu) yang tidak seimbang. Metode decision tree memiliki kinerja yang baik dalam menangani klasifikasi tepat waktu atau terlambat tetapi decision tree memiliki kelemahan dalam derajat yang tinggi dari ketidakseimbangan kelas (*class imbalance*). Untuk mengatasi masalah tersebut dapat dilakukan dengan sebuah metode yang dapat menyeimbangkan kelas dan meningkatkan akurasi. Adaboost salah satu metode *boosting* yang mampu menyeimbangkan kelas dengan memberikan bobot pada tingkat *error* klasifikasi yang dapat merubah distribusi data. Eksperimen dilakukan dengan menerapkan metode adaboost pada decision tree (DT) untuk mendapatkan hasil yang optimal dan tingkat akurasi yang baik. Hasil ekperimen yang diperoleh dari metode decision tree untuk akurasi sebesar 87,18%, AUC sebesar 0,864, dan RMSE sebesar 0,320, sedangkan hasil dari decision tree dengan adaboost (DTBoost) untuk akurasi sebesar 90,45%, AUC sebesar 0,951, dan RMSE sebesar 0,273, maka dapat disimpulkan dalam penentuan kelulusan mahasiswa dengan metode decision tree dan adaboost terbukti mampu menyelesaikan masalah ketidakseimbangan kelas dan meningkatkan akurasi yang tinggi dan dapat menurunkan tingkat *error* klasifikasi.

Keywords: kelulusan, ketidakseimbangan kelas, decision tree, adaboost

1. PENDAHULUAN

Angka partisipasi mahasiswa di setiap tahun akademik terjadi peningkatan, maka daya tampung dan tingkat kelulusan pun perlu diperhatikan dan menjadi bagian terpenting untuk evaluasi penentuan kelulusan dan dapat dijadikan sebagai bahan pendukung dalam pengambilan keputusan.

Metode klasifikasi banyak digunakan oleh para peneliti seperti Decision Tree (DT) untuk prediksi kelulusan (Undavia, Dolia, & Shah, 2013), Artificial Neural Networks (ANNs) untuk prediksi hasil kelulusan (Karamouzis & Vrettos, 2008), model klasifikasi dengan pemberian bobot menggunakan Algoritma Genetika (GA) (Minaei-Bidgoli, Kashy, Kortemeyer, & Punch, 2013).

Neural network (NN) memiliki kelebihan pada prediksi non-linier, kuat pada *parallel processing* dan kemampuan untuk mentoleransi kesalahan, tetapi memiliki kelemahan pada perlunya data training yang besar, *over-fitting*, rendahnya konvergensi, dan sifatnya yang *local optimum* (Capparuccia, Leone, & Marchitto, 2007). Decision tree (DT) dapat memecahkan masalah neural network yaitu menangani *over-fitting*, menangani atribut yang kontinu, memilih yang tepat untuk *attribute selection*, menangani *training data* dengan nilai atribut yang hilang, dan meningkat efisiensi komputasi (Quinlan, 1993), pada umumnya tingkat keberhasilan dari *decision tree* difokuskan pada dataset yang relatif seimbang (Cieslak, Hoens, Chawla, & Kegelmeyer, 2012), tetapi decision tree memiliki kelemahan misalnya *entropy* dan *gini* ketika dataset memiliki derajat yang tinggi dari *class imbalance* (Cieslak, Hoens, Chawla, & Kegelmeyer, 2012).

Distribusi *class imbalance* dari sebuah *dataset* yang telah menimbulkan kesulitan yang serius pada sebagian besar algoritma pembelajaran *classifier*, yang mengasumsikan bahwa distribusi yang relatif seimbang (Sun, Kamel, Wong, & Wang, 2007). Distribusi *class imbalance* dapat ditandai sebagai sesuatu yang memiliki lebih banyak kasus dari beberapa *class* yang lain, masalah keseimbangan adalah salah satu dimana satu *class* diwakili oleh sampel yang besar, sedangkan yang lainnya hanya diwakili oleh beberapa sampel (Sun, Kamel, Wong, & Wang, 2007), pada umumnya klasifikasi standar memiliki kinerja yang buruk pada dataset *imbalance* karena mereka dirancang untuk menjeneralisasi dari data *training* dan hasil dari hipotesis yang paling sederhana adalah yang paling sesuai dengan data.

Penanggulangan *class imbalance* dalam pendistribusian secara signifikan dapat dibantu oleh metode sampling (Hulse & Khoshgoftaar, 2009). Salah satu kebutuhan untuk memodifikasi distribusi data, dikondisikan pada fungsi evaluasi. *Re-sampling*, dengan mengembangkan *class* minoritas (positif) atau mengempiskan *class* mayoritas (negatif) dari sebuah dataset yang diberikan, telah menjadi standar *de-facto* untuk mengatasi *class imbalance* dalam berbagai domain (Chawla, Cieslak, Hall, & Joshi, 2008). Beberapa teknik untuk mengatasi *class imbalance* seperti *over-sampling* cenderung mengurangi jumlah pemangkasan yang terjadi, sedangkan *under-sampling* sering membuat pemangkasan yang tidak perlu (Drummond & Holte, 2003), Drummond dan Holte dengan menggunakan metode C4.5 menemukan, bahwa *under-sampling* mayoritas adalah lebih efektif dalam menangani masalah *class imbalance* dan *over-sampling* minoritas sering menghasilkan sedikit atau tidak ada perubahan dalam kinerja. Selain itu juga mereka mencari bahwa *over-sampling* dapat dibuat *cost-sensitive* jika pemangkasan dan parameter berhenti diawal ditetapkan secara proporsional dengan jumlah lebih dari *over-sampling* yang dilakukan (Drummond & Holte, 2003).

Maka dari itu *decision tree* dengan kasus *class imbalance* diperlukan metode yang dapat mengatasi masalah tersebut untuk meningkatkan kinerja klasifikasi *decision tree* agar dapat menghasilkan kinerja yang lebih baik. Algoritma *boosting* adalah algoritma iteratif yang memberikan bobot yang berbeda

pada distribusi training data pada setiap iterasi. Setiap iterasi *boosting* menambahkan bobot pada contoh-contoh kesalahan klasifikasi dan menurunkan bobot pada contoh klasifikasi yang benar, sehingga secara efektif dapat merubah distribusi pada data training (Kotsiantis, Kanellopoulos, & Pintelas, 2006, hal. 25-36). Metode *Boosting* (AdaBoost) yang diusulkan (Kotsiantis & Pintelas, 2009, hal. 123) dengan *selective costing ensemble* dapat menjadi solusi yang lebih efektif untuk masalah *class imbalance* dan memungkin meningkatkan identifikasi dari *class* minoritas yang sulit serta menjaga kemampuan klasifikasi dari *class* mayoritas. Karena adaboost metode pembelejaran *ensamble* yang dapat mengurangi varian, hal ini terjadi karena efek bias rata-rata *ensamble* untuk mengurangi varian dari satu set pengkalsifikasian. Bias dapat dicirikan sebagai ukuran kemampuan untuk menjeneralisasi benar untuk satu set tes. Selain itu pendekatan untuk mengatasi masalah tersebut dapat dilakukan dengan beberapa metode (Weiss, McCarthy, & Zabar, 2007) yaitu *over-sampling*, *under-sampling*, dan *cost-sensitive*.

Pada penelitian ini yang akan dilakukan adalah penerapan adaBoost untuk penyelesaian *class imbalance* pada *decision tree* untuk penentuan kelulusan, sehingga dapat menghasilkan kinerja yang baik pada dataset yang tidak seimbang.

2. PENELITIAN TERKAIT

Penelitian tentang prediksi kelulusan telah banyak dilakukan dan telah dipublikasikan. Untuk melakukan penelitian ini perlu ada kajian terhadap penelitian yang terkait sebelumnya agar dapat mengetahui metode apa saja yang digunakan, data seperti apa yang diproses, dan model seperti apa yang dihasilkan.

Undavia, Dolia, dan Shah (2013) dalam penelitiannya pada lembaga pendidikan tinggi yang mengalami rendahnya persentase dari hasil penempatan dan minat mahasiswa menggunakan *decision tree* pada sistem pendukung keputusan untuk memprediksi pasca kelulusan bagi mahasiswa berdasarkan prestasi akademik pada data historis. Hasil akurasi sebesar 67,1875% dan Kappa yang diperoleh 0,1896. Jumlah Confusion Matrix 86 (73+13) record klasifikasi benar untuk MBA dan MCA, sedangkan 42 (32+10) record diklasifikasikan salah untuk kedua Program PG.

Ramesh, Parkavi, dan Ramar (2013) dalam penelitiannya mengidentifikasi faktor-faktor yang mempengaruhi prestasi siswa dalam ujian akhir dan mencari tahu algoritma mana yang cocok untuk memprediksi *grade* dari siswa sehingga dapat memberikan secara tepat waktu dan memberikan peringatan bagi siswa mereka yang beresiko. Hasil yang diperoleh dari pengujian hipotesis menunjukan bahwa jenis sekolah tidak mempengaruhi prestasi siswa dan kedudukan orang tua memiliki peran utama dalam memprediksi *grade.* Algoritma klasifikasi terbaik yang digunakan dalam studi tersebut yaitu Multilayer Perception. Pencarian atribut terbaik dengan menggunakan metode "Ranker" dan mendapatkan 10 atribut teratas yang dipilih dari 27 atribut, sebanyak 500 record diambil untuk dijadikan analisis. Hasil akurasi ternyata Multilayer Perception memiliki akurasi yang terbaik sebesar 72,38%.

Marquez-Vera, Cano, Romero, dan Ventura (2013) dalam penelitiannya memprediksi kegagalan siswa di Sekolah dengan mengukur faktor yang dapat mempengaruhi rendahnya prestasi siswa dan sifat dari dataset yang tidak seimbang (*imbalance*). Algoritma yang diusulkan yaitu *genetic programming* dan pendekatan data mining yang berbeda. Data yang digunakan sekitar 670 siswa tingkat atas dari sekolah Zacatecas – Meksiko.

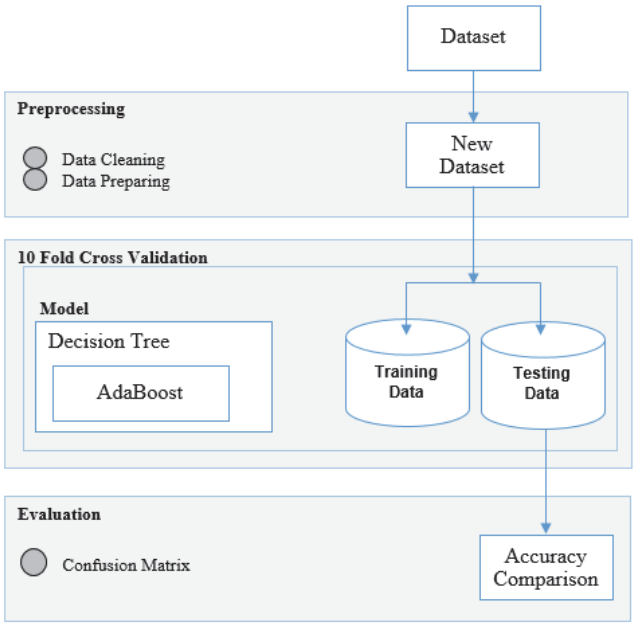
Sebagian besar kasus data yang telah dilakukan untuk klasifikasi siswa gagal atau dropout adalah tidak seimbang, yang berarti bahwa hanya sebagian kecil siswa gagal dan sebagian besar lulus. Metode yang diusulkan mampu memecahkan masalah dalam dunia pendidikan, seperti menggunakan klasifikasi *cost-sensitive* dan *resampling* dari dataset asli.

Thammasiri, Delen, Meesad, dan Kasap (2014) dalam penelitiannya membandingkan perbedaan teknik data yang tidak seimbang untuk meningkatkan akurasi prediksi dalam *class* minoritas tetapi tetap mempertahankan kepuasan kinerja klasifikasi secara keseluruhan. Secara khusus Thammasiri menggunakan tiga metode pengujian *balancing* yaitu teknik *over-sampling*, *under-sampling* dan *synthetic minority over-sampling* (SMOTE) beserta empat metode klasifikasi yang populer (LR, DT, NN, dan SVM). Hasil penelitian menunjukan bahwa SVM dengan SMOTE mencapai kinerja terbaik dengan tingkat akurasi mencapai 90.24% pada sampel *10-fold holdout*.

Pada penelitian ini berbeda dengan penelitian yang telah ada, untuk klasifikasi menggunakan decision tree, sedangkan untuk mengatasi masalah ketidakseimbangan kelas menggunakan metode adaboost untuk menyelesaikan masalah data atau kelas yang sifatnya tidak seimbang (*imbalance*).

3. METODE YANG DIUSULKAN

Metode yang diusulkan dalam penelitian ini yaitu untuk meningkatkan kinerja algoritma decision tree dengan metode adaboost yang dapat menangani ketidakseimbangan kelas pada klasifikasi dataset kelulusan. Sedangkan untuk validasi menggunakan *10-fold cross validation*. Hasil pengukuran dengan analisa menggunakan t-Test. Model kerangka pemikiran metode yang diusulkan ditunjukan pada Gambar 1.



Gambar 1. Kerangka Pemikiran Model yang diusulkan

Pada Gambar 1 dalam pengolahan awal, data yang sudah didapat dibersihkan dan pilah, sehingga menjadi sebuah dataset baru untuk *training* dan testing dari atribut yang sudah ditentukan. Setelah itu dimasukan kedalam *classifier* dengan

metode algoritma decision tree, kemudian dataset yang tidak seimbang diselesaikan oleh *boosting* (adaBoost). Pada dasarnya, metode *boosting* dapat meningkatkan ketelitian dalam proses klasifikasi dan prediksi dengan cara membangkitkan kombinasi dari suatu model, tetapi hasil klasifikasi atau prediksi yang dipilih adalah model yang memiliki nilai bobot paling besar. Jadi, setiap model yang dibangkitkan memiliki atribut berupa nilai bobot. Dataset yang telah seimbang akan divalidasi dengan menggunakan *10-fold cross validation*. Hasil dari validasi akan menghasilkan data yang diukur yaitu AUC dan *Accuracy*.

Zhou dan Yu (2009) menjelaskan teknik pembobotan pada algoritma adaboost sebagai berikut:

Input:
Dataset (D_t) = (x_1, y_1), …, (x_m,y_m);
Weak Lern (*L*)
T menyatakan jumlah iterasi
Proses:
Inisialisasi nilai bobot
$D_1(i) = \frac{1}{m}$ untuk i = 1, …, *m*
for *t* = 1, … , *T*;
Pengujian terhadap distribusi D_t
$h_t = L(D_t)$
hitung *error* dari $\epsilon_t = Pr_{x\sim D_t, y}\, I[h_t(x) \neq y]$
if $\epsilon_t > 0.5$ then break
menentukan bobot dari h_t
$\alpha t = \frac{1}{2} ln\left(\frac{1-\epsilon t}{\epsilon t}\right)$;
$D_{t+1}(i) = \frac{D_t(i)}{Z_i}$ x $\begin{cases}\exp(-\alpha_t)\ if\ h_t(x_i) = y_i \\ \exp(\alpha_t)\ if\ h_t(x_i) \neq y_i\end{cases}$

Update distribusi, dimana *Zt* sebuah faktor normalisasi yang mengaktifkan D_{t+1} menjadi distribusi

$$\frac{Dt(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$$

end for
Output:
H(x) = $sign(\sum_{t=1}^{T} \propto_t\ h_t(x))$

Pada tahap evaluasi dilakukan dengan memperoleh hasil AUC, *Accuracy*, kemudian validasi dilakukan dengan pengujian *Root Mean Square Error* (RMSE) dengan uji t-Test untuk mengetahui apakah ada perbedaan antara metode decision tree dengan decision tree dan adaboost.

Evaluasi terhadap model mengukur akurasi dengan *confusion matrix* yang menitikberatkan pada *class* secara umum, sedangkan untuk AUC menggunakan ROC Curve dan proses dengan menggunakan *10 fold cross validation*.

Pengukuran akurasi dengan *confusion matrix* dapat dilihat pada Tabel 1.

Tabel 1. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Formulasi perhitungan yang digunakan (Gorunescu, 2011) adalah sebagai berikut:

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

$$Sensitivity = TP\ rate = \frac{TP}{TP+FN}$$

$$Specificity = TN\ rate = \frac{TN}{TN+FP}$$

$$FP\ rate = \frac{FP}{FP+TN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F\text{-}Measure = \frac{2RP}{R+P}$$

$$G\text{-}Mean = \sqrt{Sensitivity * specificity}$$

Evaluasi dengan *F-Measure*, rata-rata harmonik dari dua angka cenderung lebih dekat dengan lebih kecil dari dua, oleh karena itu nilai *F-Measure* yang tinggi dapat memastikan bahwa kedua *recall* (*sensitivity*) dan presisi yang cukup tinggi. Jika hanya kinerja kelas positif dianggap sebagai dua langkah penting yaitu *TP rate* dan *Positive Predictive Value* (*PP value*). *PP value* didefinisikan sebagai presisi yang menunjukan presentasi objek yang relevan yang didefinisikan untuk retrieval. Dalam pencarian informasi *TP rate* didefinisikan sebagai *recall* yang menunjukan presentasi dari objek yang diambil itu adalah relevan. Rata-rata harmonik adalah gabungan dari ukuran presisi dan *recall*.

Evaluasi dengan *Receiver Operating Character Curve* (*ROC Curve*), secara teknis menggambarkan graifk dua dimensi, dimana tingkat *True Positive* (*TP*) terletak pada garis sumbu *Y*, sedangkan untuk *False Positive* (*FP*) terletak pada garis sumbu *X*. dengan demikian *ROC* menggambarkan *trade-off* antara *TP* dan *FP*. Pencatatan dalam *ROC* dinyatakan dalam sebuah klausa yaitu semakin rendah titik kekiri (0.0), maka dianyatakan sebagai klasifikasi prediksi mendekati/menjadi negatif, sedangkan semakin keatas titik kekanan (1.1), maka dinyatakan sebagai klasifikasi prediksi mendekati/menjadi positif. Titik dengan nilai 1 dinyatakan sebagai tingkat *True Positif* (*TP*), sedangkan titik dengan nilai 0 dinyatakan sebagai tingkat *False Positive* (*FP*). Pada titik (0.1) merupakan klasifikasi prediksi adalah sempurna karena semua kasus baik positif maupun negatif dinyakatakan dengan benar (*True*). Sedangkan untuk (1.0) klasifikasi prediksi semuanya dinyatakan sebagai tidak benar (*False*).

Dalam pengklasifikasian keakuratan dari tes diagnostik menggunakan *Area Under Curve* (*AUC*) (Gorunescu, 2011, hal. 325-326) sebuah sistem nilai yang disajikan pada Tabel 2.

Tabel 2. Nilai AUC dan Keterangan

AUC	Keterangan
0.90 - 1.00	*excellent classification*
0.80 - 0.90	*good classification*
0.70 - 0.80	*fair classification*
0.60 - 0.70	*poor classification*
< 0.60	*failure*

Evaluasi dengan *Root mean square error* (RMSE) adalah sebuah metode konvensional yang digunakan untuk menghitung rata-rata ukuran dari sebuah deviasi dan disebut juga sebagai perbedaan antara nilai aktual dengan nilai prediksi (Barreto & Howland, 2006). RMSE adalah estimasi dari sebuah sampel peta dan referensi poin (Congalton & Green, 2009). Jadi, *root mean square error* (RMSE) atau disebut juga sebagai root *mean square deviation* (RMSD) suatu ukuran yang digunakan dari perbedaan antara nilai-nilai yang diprediksi oleh sebuah model dengan nilai-nilai aktual.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i' - y_i)^2}{n}}$$

Keterangan formulasi RMSE:
y' : nilai aktual
y : nilai prediksi
n : jumlah sampel data
i : terasi

4. HASIL EKSPERIMEN

Eksperimen dilakukan dengan menggunakan sebuah platform komputer berbasis Intel Core i3-3220 @3.30GHz (4 CPUs), RAM 4GB, dan sistem operasi Microsoft Windows 8.1 Profesional 64-bit. Sedangkan lingkungan pengembangan aplikasi dengan bahasa pemrograman Java Netbeans IDE 8.0 dan Rapidminer 6.0, untuk analisis hasil menggunakan aplikasi Excel Data Anlysis.

Data kelulusan mahasiswa yang bisa digunakan sebagai dataset yaitu dengan jumlah mahasiswa 429 record dengan status lulus. Dari jumlah tersebut memiliki 15 atribut dengan status klasifikasi tepat waktu sebesar 119 record (27,74%) dan tidak tepat waktu atau terlambat sebesar 310 record (72,26%). Karakterisitik dari dataset universitas pamulang dapat dilihat pada Tabel 3.

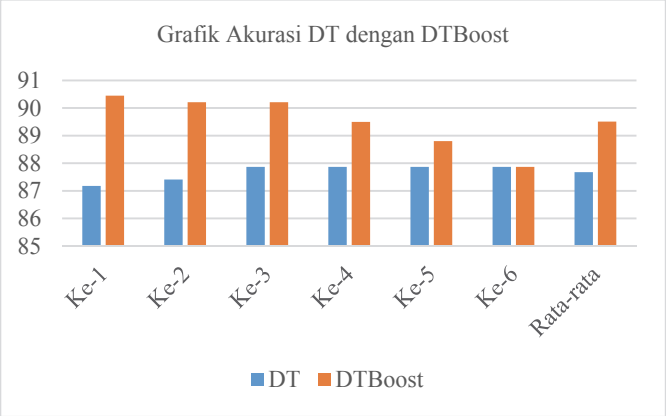
Tabel 3. Karakterisitik Dataset Universitas Pamulang

No	Atribut	Tipe Data	Keterangan
1	NIM	Nominal	Nomor Induk Mahasiswa (ID)
2	Prodi	Nominal	Program Studi
3	JK	Nominal	Jenis Kelamin
4	Shift	Nominal	Shift waktu perkuliahan
5	Usia	Numeric	Usia Mahasiswa
6	IPS-1	Numeric	Indek Prestasi Semester – 1
7	IPS-2	Numeric	Indek Prestasi Semester – 2
8	IPS-3	Numeric	Indek Prestasi Semester – 3
9	IPS-4	Numeric	Indek Prestasi Semester – 4
10	IPS-5	Numeric	Indek Prestasi Semester – 5
11	IPS-6	Numeric	Indek Prestasi Semester – 6
12	IPS-7	Numeric	Indek Prestasi Semester – 7
13	IPS-8	Numeric	Indek Prestasi Semester – 8
14	IPK	Numeric	Indek Prestasi Kumulatif
15	Status	Nominal	Tepat Waktu atau Tidak Tepat Waktu (Label)

Dalam eksperimen ini dilakukan dengan nilai parameter decision tree: *criterion*, *minimal size for split*, *minimal leaf size*, *minmal gain*, *maximal depth*, *confidence* dan *split vaildation* dengan parameter *split*, *split ration*, *stratified sampling* untuk mendapatkan nilai AUC, *Accuracy*, dan RMSE. Hasil eksperimen dapat dilihat pada Tabel 4 sampai dengan Tabel 8.

Tabel 4. Perbandingan Akurasi

Eksperimen	DT	DTBoost
Ke-1	87,18	90,45
Ke-2	87,41	90,21
Ke-3	87,87	90,21
Ke-4	87,87	89,50
Ke-5	87,87	88,80
Ke-6	87,87	87,87
Rata-rata	**87,68**	**89,51**



Gambar 2. Grafik Perbandingan Akurasi

Tabel 5. Confusion Matrix Model DT

	True Terlambat	True Tepat	Class Precision
Pred. Terlambat	274	19	93.52%
Pred. Tepat	36	100	73,75%
Class Recall	88,39%	84,03%	

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

$$Accuracy = \frac{(274+100)}{(274+19+100+36)}$$

$$Accuracy = 0{,}8718 = 87{,}18\%$$

Dari jumlah data sebanyak 429 klasifikasi kelas dengan status terlambat sebesar 310 record dan status tepat sebesar 119 record. Data diprediksi yang sesuai dengan status terlambat sejumlah 274, data yang diprediksi terlambat tetapi kenyataannya tepat sejumlah 19, data yang diprediksi tepat tetapi kenyataannya terlambat sejumlah 36, dan sedangkan data yang diprediksi tepat dan sesuai sejumlah 100.

Tabel 6. Confusion Matrix Eksperimen DTBoost

	True Terlambat	True Tepat	Class Precision
Pred. Terlambat	291	22	92,97%
Pred. Tepat	19	97	83,62%
Class Recall	93,87%	81,51%	

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

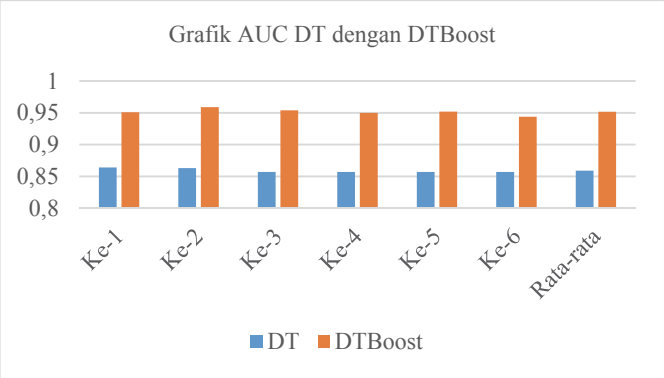
$$Accuracy = \frac{(291+97)}{(291+22+97+19)}$$

$$Accuracy = 0{,}9045 = 90{,}45\%$$

Dari jumlah data sebanyak 429 klasifikasi kelas dengan status terlambat sebesar 310 record dan status tepat sebesar 119 record. Data diprediksi yang sesuai dengan status terlambat sejumlah 291, data yang diprediksi terlambat tetapi kenyataannya tepat sejumlah 22, data yang diprediksi tepat tetapi kenyataannya terlambat sejumlah 19, dan sedangkan data yang diprediksi tepat dan sesuai sejumlah 97.

Tabel 7. Perbandingan Area Under Curve (AUC)

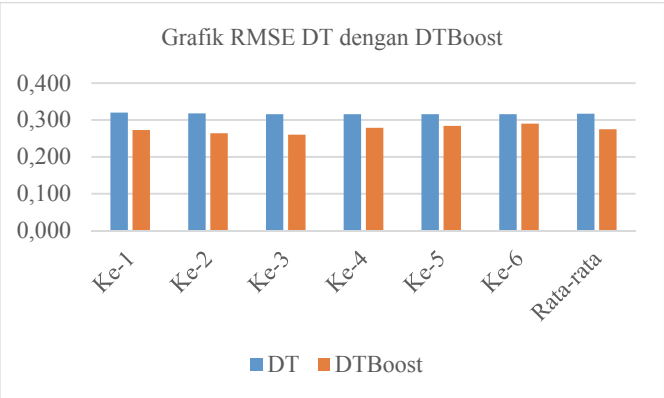
Eksperimen	DT	DTBoost
Ke-1	0,864	0,951
Ke-2	0,863	0,959
Ke-3	0,857	0,954
Ke-4	0,857	0,950
Ke-5	0,857	0,952
Ke-6	0,857	0,944
Rata-Rata	**0,859**	**0,952**



Gambar 3. Grafik Perbandingan AUC

Tabel 8. Perbandingan RMSE

Eksperimen	DT	DTBoost
Ke-1	0,320	0,273
Ke-2	0,318	0,264
Ke-3	0,316	0,260
Ke-4	0,316	0,279
Ke-5	0,316	0,284
Ke-6	0,316	0,290
Rata-rata	**0,317**	**0,275**



Gambar 4. Grafik RMSE DT dengan DTBoost

Pada penelitian ini dilakukan pengujian hipotesis dengan uji *paired sample t-Test* decision tree (DT) dengan decision tree dan adaboost (DTBoost). T-Test adalah hubungan antara variabel respon dengan varibel prediktor (Larose, 2007). Hipotesis nol (H_0) menyatakan bahwa tidak ada hubungan yang linier antara varibel-variabel, sedangkan hipotesis alternatif (Ha) menyatakan bahwa adanya hubungan antara variabel-variabel. H_0 merupakan tidak ada perbedaan antara DT dan DTBoost, Ha merupakan ada perbedaan antara DT dan DTBoost. Pada *paired sample t-Test* dengan *root mean square error* (RMSE) yang terdiri dari variabel DT dan variabel DTBoost dapat dilihat pada Tabel 9.

Tabel 9. *Paired Samples Test* RMSE DT dengan DTBoost

	DT	*DTBoost*
Mean	0,317	0,275
Variance	2,8E-06	0,0001344
Observations	6	6
Pearson Correlation	-0,309294787	
Hypothesized Mean Difference	0	
df	5	
t Stat	8,42249	
P	**0,000193471**	

Pada Tabel 9 menunjukan hasil dari *paired sample test* RMSE DT dengan DTBoost, bahwa untuk nilai uji t memiliki aturan apabila *P-value* < 0,05 terdapat perbedaan pada taraf signifikan yaitu 5%, dan apabila *P-value* > 0,05, maka tidak ada perbedaan antara sebelum dan sesudah. Hasil yang didapat dari nilai uji t untuk *P-value* sebesar 0,000193471 < 0,05 yang berarti bahwa H_0 ditolak atau Ha diterima, adanya perbedaan yang signifikan antara DT dan DTBoost.

Metode decision tree dengan adaboost menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan menggunakan decision tree versi standar. Hal tersebut seperti dikatakan oleh Quinlan, bahwa adaboost dapat memberikan keuntungan, lebih efektif dan akurat dalam pengklasifikasian (Quinlan, Bagging, Boosting, and C4.5, 1996).

5. KESIMPULAN

Penelitian dengan menerapkan metode adaboost untuk penyelesaian ketidakseimbangan kelas (*class imbalance*) pada penentuan kelulusan mahasiswa dengan metode decision tree, eksperimen telah dilakukan untuk mendapatkan sebuah model arsitektur yang optimal dan mendapatkan hasil estimasi yang akurat. Hasil pengujian diatas dapat disimpulkan bahwa metode adaboost sebagai metode boosting terbukti efektif dalam penyelesaian ketidakseimbangan kelas pada penentuan kelulusan mahasiswa dengan metode decision tree.

REFERENSI

Barreto, H., & Howland, F. M. (2006). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft ExceIntroductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel.* New York: Cambridge University Press.

Capparuccia, R., Leone, R. D., & Marchitto, E. (2007). Integrating support vector machines and neural networks. *Neural Networks*, 590-597.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, 225-252.

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 136-158.

Congalton, R. G., & Green, K. (2009). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Second Edition (Mapping Science).* Boca Raton: CRC Press.

Drummond, C., & Holte, R. C. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. *Institute for Information Technology,*

National Research Council (pp. 1-8). Canada, Ottawa, Ontario: Department of Computing Science, University of Alberta.

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

Hulse, J. V., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. *Elsevier*, 1513-1542.

Karamouzis, S. T., & Vrettos, A. (2008). An Artificial Neural Network for Predicting Student Graduation Outcomes. *WCECS (World Congress on Engineering and Computer Science)*, 991-994.

Kotsiantis, S. B., & Pintelas, P. E. (2009). Selective costing ensemble for handling imbalanced data sets. *International Journal of Hybrid Intelligent Systems*, 123-133.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 25-36.

Larose, D. T. (2007). *Data Mining Methods and Models.* Hoboken, New Jersey: A John Wley & Sons, Inc Publication.

Marquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, 315-330.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2013). Predicting Student Performance: An Application Of Data Mining Methods With The Educational Web-Based System Lon-Capa. *IEEE (Institute of Electrical and Electronics Engineers)*.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Quinlan, J. R. (1996). Bagging, Boosting, and C4.5. *AAAI'96 Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1* (pp. 725-730). Australia: ACM Digital Library.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *International Journal of Computer Applications*, 35-39.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition Society*, 3358–3378.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications: An International Journal*, 321-330 .

Undavia, J. N., Dolia, P. M., & Shah, N. P. (2013). Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment. *International Journal of Computer Applications*.

Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? *DMIN*, 35-41.

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

Zhou, Z.-H., & Yu, Y. (2009). *The Top Ten Algorithms in Data Mining.* (X. Wu, & V. Kumar, Eds.) Chapman & Hall/CRC.

BIOGRAFI PENULIS

Achmad Bisri. Memperoleh gelar Sarjana Komputer (S.Kom) dibidang Teknik Informatika dari STMIK Banten Jaya, Serang-Banten, gelar Magister Komputer (M.Kom) dibidang Software Engineering (Rekayasa Perangakat Lunak) dari STMIK Eresha, Jakarta. Dia saat ini sebagai staf pengajar di Universitas Pamulang (Unpam), Tangerang Selatan. Minat Penelitiannya saat ini meliputi software engineering (rekayasa perangkat lunak) dan mechine learning.

Romi Satria Wahono. Memperoleh gelar B.Eng. dan M.Eng. di bidang ilmu komputer dari Saitama University, Jepang dan gelar Ph.D. di bidang Software Engineering dari Universiti Teknikal Malaysia Melaka. Saat ini sebagai pengajar dan peneliti pada program Pascasarjana Ilmu Komputer di Universitas Dian Nuswantoro, Indonesia. Juga merupakan pendiri dan CEO PT Brainmatics Cipta Informatika, sebuah perusahaan pengembangan perangkat lunak di Indonesia. Minat penelitiannya saat ini meliputi rekayasa perangkat lunak dan machine learning. Anggota profesional dari asosiasi ilmiah ACM, PMI dan IEEE Computer Society.