Integrasi *Gradient Boosted Trees* dengan SMOTE dan *Bagging* untuk Deteksi Kelulusan Mahasiswa

Achmad Bisri[1], Rinna Rachmatika[2]

***Abstract*—Education has an important role in life. Pamulang University is a university which provides education at affordable cost. However, based on student academic performance data, there is imbalance in class between the number of students who graduate on time and students who can not graduate on time, on various study programs. In this paper, an implementation of SMOTE and bagging techniques was conducted on the Gradient Boosted Trees (GBT) classification method for handling the class imbalance problem. The proposed method is able to provide significant results with an accuracy of 80.57% and an AUC of 0.858, in the category of good classification.**

***Intisari*—Pendidikan memiliki peranan penting dalam membangun dan mencerdaskan di dalam kehidupan. Universitas Pamulang adalah sebuah pendidikan tinggi yang memberikan kemudahan dengan biaya terjangkau. Namun, berdasarkan data prestasi akademik mahasiswa/i, terdapat ketidakseimbangan kelas antara jumlah mahasiswa/i yang lulus tepat waktu dan tidak tepat waktu dari berbagai program studi. Dalam makalah ini, dilakukan penerapan SMOTE dan teknik *bagging* pada metode klasifikasi *Gradient Boosted Trees* (GBT) untuk mengatasi masalah ketidakseimbangan kelas. Metode yang diusulkan mampu memberikan hasil yang signifikan dengan nilai akurasi sebesar 80,57% dan nilai AUC sebesar 0,858, dalam kategori klasifikasi yang baik (*good classification*).**

***Kata Kunci*—*Gradient Boosted Trees*, SMOTE, *Bagging*, Deteksi Kelulusan, Ketidakseimbangan Kelas.**

I. PENDAHULUAN

Pendidikan memiliki peranan penting dalam kehidupan. Universitas Pamulang adalah sebuah pendidikan tinggi yang hingga saat ini terus mengembangkan kualitas, sumber daya pendidik dan staf kependidikan, infrastruktur, sarana, dan prasarana untuk mewujudkan pendidikan dengan biaya terjangkau oleh seluruh lapisan masyarakat tanpa mengabaikan kualitas.

Berdasarkan data yang diperoleh dari pangkalan data pendidikan tinggi (https://forlap.ristekdikti.go.id), saat ini Universitas Pamulang memiliki tujuh belas program studi yang terbagi pada tiga jenjang, yaitu dua program studi pada jenjang strata dua (S2), tiga belas program studi pada jenjang strata satu (S1), dan dua program studi pada jenjang diploma tiga (D3), dengan jumlah mahasiswa/i secara keseluruhan sebesar 72.571 mahasiswa/i pada pelaporan tahun 2018/2019.

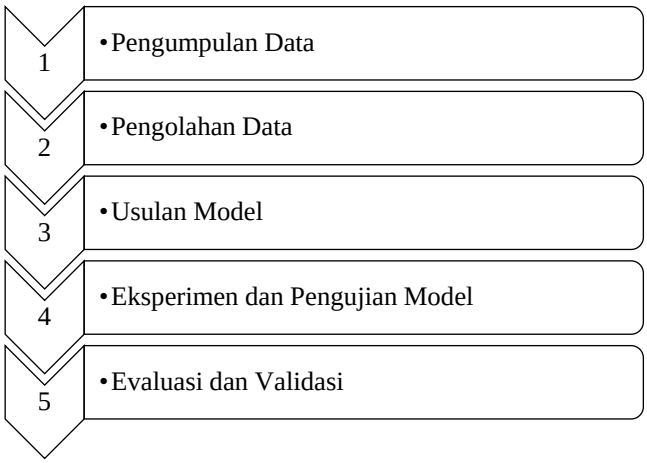
Jumlah mahasiswa/i yang besar dalam suatu pendidikan tinggi perlu mendapatkan perhatian khusus, dari penerimaan mahasiswa/i baru sebagai *input* hingga lulusan mahasiswa/i sebagai *output*. Saat ini, Universitas Pamulang memiliki jumlah kelulusan mahasiswa/i tepat waktu lebih sedikit dibandingkan dengan jumlah kelulusan mahasiswa/i tidak tepat waktu. Di sisi lain, kelulusan tidak tepat waktu dapat meningkat dengan bertambahnya jumlah mahasiswa/i yang melakukan cuti perkuliahan. Hal tersebut mengakibatkan ketidakseimbangan antara jumlah yang lulus tepat waktu dan jumlah yang lulus tidak tepat waktu, sehingga deteksi kelulusan mahasiswa/i penting untuk dilakukan, sebagai sarana pengambilan keputusan dan kebijakan lain seperti kebutuhan rekrutmen mahasiswa/i baru dan penilaian akreditasi.

Berbagai teknik *machine learning* pada *data mining* yang digunakan oleh para peneliti untuk melakukan deteksi kelulusan tepat waktu dan tidak tepat waktu telah dilakukan. Salah satunya adalah prediksi kelulusan menggunakan metode *decision tree* dengan *AdaBoost* untuk penentuan kelulusan mahasiswa/i tepat waktu dan tidak tepat waktu dalam penanganan ketidakseimbangan kelas [1]. Prediksi ini menggunakan *dataset* bersifat *private* yang terdiri atas 429 *record* dan lima belas atribut dengan status kelulusan mahasiswa/i tepat waktu sebesar 119 *record* dan tidak tepat waktu sebesar 310 *record* [1]. Prediksi kelulusan mahasiswa/i menggunakan *Artificial Neural Network* (ANN) juga telah dilakukan, dengan data masukan dari tahun akademik 2009/2010, dengan kelulusan tahun 2013 sebesar 193 mahasiswa/i dari program studi Teknik Informatika [2]. Contoh penelitian lainnya adalah pada seleksi calon mahasiswa/i baru dengan menggunakan perbandingan berbagai metode klasifikasi dan mendapatkan hasil yang terbaik pada metode *Support Vector Machine* (SVM) [3]. Data yang digunakan berasal dari penerimaan mahasiswa/i baru dengan enam atribut dari data mahasiswa yang sudah lulus dan juga yang masa studinya sudah delapan semester, dengan status kelulusan tepat waktu berjumlah 64 dan tidak tepat waktu berjumlah 107 dari keseluruhan data berjumlah 171 kelulusan mahasiswa/i [3].

Namun, berdasarkan data kelulusan pada umumnya yang digunakan pada penentuan kelulusan mahasiswa/i, terdapat masalah pada ketidakseimbangan kelas yang disebabkan distribusi data yang tidak seimbang antara data kelas kelulusan tepat waktu dan data kelas kelulusan tidak tepat waktu. Jumlah data kelas kelulusan yang tepat waktu lebih sedikit dibandingkan data kelas kelulusan yang tidak tepat waktu, sehingga dapat menyebabkan tingkat akurasi cenderung pada jumlah data kelas yang mayoritas.

Penanganan ketidakseimbangan kelas secara umum dapat dilakukan dengan pendekatan pada tingkatan data dan tingkatan secara algoritmis. Pada tingkatan data, penanganan dapat dilakukan dengan berbagai teknik *sampling*, seperti *random under-sampling* maupun *random over-sampling*,

[1,2] *Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pamulang, Jalan Surya Kencana No. 1, Pamulang, Tangerang Selatan 15417 INDONESIA (telp.: 021-7412566; e-mail: achmadbisri@unpam.ac.id, rinnarachmatika@unpam.ac.id)*



Gbr. 1 Tahapan penelitian.

sedangkan pada tingkatan algoritmis, dapat dilakukan dengan beberapa kategori seperti klasifikasi pembelajaran satu kelas, pembelajaran *cost-sensitive*, pembelajaran *ensemble*, dan pembelajaran *hybrid* [4], [5]. Pendekatan dengan pembelajaran *ensemble* pada ketidakseimbangan kelas dapat dilakukan dengan berbagai teknik, seperti *bagging* dan *boosting* [6].

Pada makalah ini, dilakukan pendekatan pada tingkatan data dan tingkatan algoritmis. *Synthetic Minority Over-Sampling Technique* (SMOTE) digunakan untuk melakukan *sampling* dengan membuat sampel data kelas minoritas dengan interpolasi beberapa sampel [7]. Metode *Gradient Boosted Trees* (GBT) dipilih sebagai metode atau algoritme klasifikasi yang andal menangani data numerik, nominal, maupun data yang hilang (*missing data*) dalam membangun sebuah model klasifikasi dibandingkan dengan metode klasifikasi lain [8]. Teknik *bagging* dipilih sebagai metode *ensemble* yang dapat mereduksi tingkat kesalahan pada model klasifikasi (*misclassification*) dalam ketidakseimbangan kelas dan dapat meningkatkan stabilitas kinerja dari model klasifikasi yang dibangun [9].

II. METODE

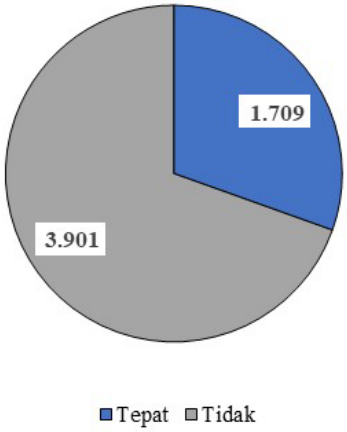
Penelitian dalam makalah ini secara sistematis dibagi ke dalam tahapan-tahapan penelitian yang terdiri atas pengumpulan data, pengolahan data, usulan model, eksperimen, dan pengujian model serta evaluasi dan validasi, seperti ditunjukkan pada Gbr. 1.

A. Pengumpulan Data

Pada tahapan ini, pengumpulan data dilakukan terhadap *raw* data yang bersifat khusus yang diperoleh dari pusat data akademik Universitas Pamulang. Berdasarkan data empiris dari tahun 2013 dengan berbagai program studi dari beberapa fakultas, data berjumlah ribuan sampel dari kelulusan mahasiswa/i tahun 2017 dan 2018.

B. Pengolahan Data

Pengolahan data dilakukan dari hasil pengumpulan data dengan melakukan pembersihan data (*data cleaning*) untuk mengatasi masalah data, seperti anomali, nilai yang hilang, redundansi data, dan data yang tidak sesuai. Setelah itu, data



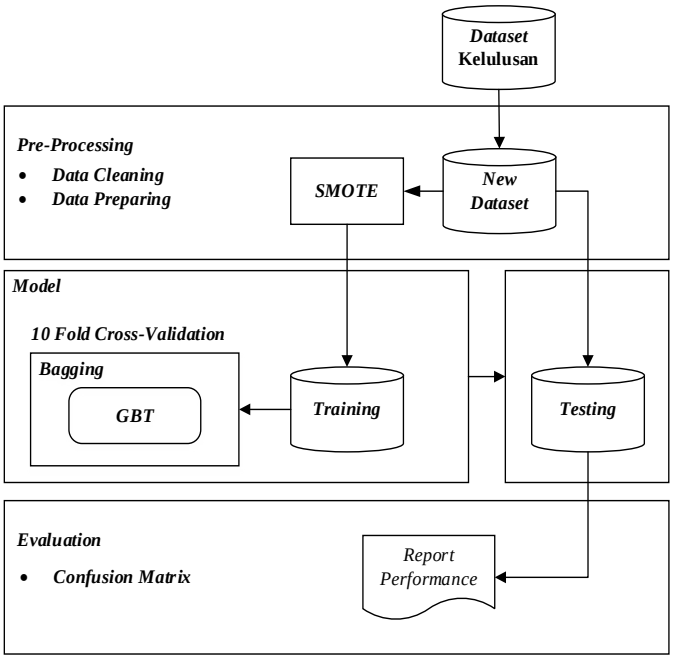
Gbr. 2 Ketidakseimbangan kelas pada data kelulusan.

TABEL I
KARAKTERISTIK *DATASET* KELULUSAN

No.	Atribut	Tipe Data	Keterangan
1	Usia	*Integer*	Usia Mahasiswa
2	JK	*Binominal*	Jenis Kelamin (L / P)
3	Prodi	*Polynominal*	Program Studi
4	Reg.	*Polynominal*	Jenis Reguler (A, B, dan C)
5	IPS_1	*Real*	IP Semester 1
6	IPS_2	*Real*	IP Semester 2
7	IPS_3	*Real*	IP Semester 3
8	IPS_4	*Real*	IP Semester 4
9	IPS_5	*Real*	IP Semester 5
10	IPS_6	*Real*	IP Semester 6
11	IPS_7	*Real*	IP Semester 7
12	Status	*Binominal*	*Class* (Tepat atau Tidak)

dipilah dan dikelompokkan sesuai jenis dan fungsi untuk distribusi ke dalam data *training* dan *testing*, yang akan diterapkan pada model klasifikasi. Data hasil pengolahan diperoleh sebesar 5.610 *instance*, yang terdiri atas 1.709 kelulusan yang tepat waktu dan 3.901 kelulusan yang tidak tepat waktu. Gbr. 2 menunjukkan data ketidakseimbangan kelas antara kelulusan tepat waktu dan kelulusan yang tidak tepat waktu. Sedangkan atribut *dataset* yang ditunjukkan pada Tabel I merupakan karakteristik *dataset* kelulusan yang sudah diolah, sebagian besar adalah atribut yang digunakan pada model klasifikasi, yaitu Indeks Prestasi Semester (IPS) dari semester satu sampai dengan semester tujuh.

Tabel I merupakan bentuk karakteristik *dataset* kelulusan yang digunakan dalam membangun model klasifikasi pada deteksi kelulusan mahasiswa/i yang terdiri atas usia; jenis kelamin (L/P); sembilan program studi pada jenjang strata satu, yaitu program studi Akuntansi, Ilmu Hukum, Manajemen, Sastra Indonesia, Sastra Inggris, Pendidikan Ekonomi, Pendidikan Pancasila dan Kewarganegaraan, Teknik Informatika, dan Teknik Mesin; jenis reguler yang terdiri atas reguler A, B, dan C; dan IPS yang dimulai dari semester satu hingga semester tujuh, dengan status kelulusan tepat waktu dan tidak tepat waktu sebagai *class* atau label.



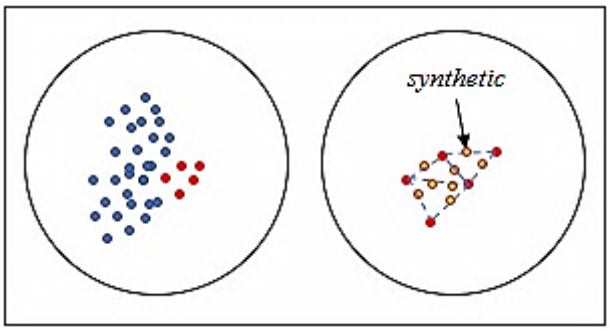
Gbr. 3 Skema usulan model.

C. Usulan Model

Pada Gbr. 3 ditunjukkan alur diagram model yang diusulkan. *Dataset* berisi 5.610 *instance* dibagi ke dalam data *training* dan data *testing*. Pembagian dilakukan dengan menggunakan teknik *shuffled sampling* ke dalam tiga bagian, yaitu 70:30, 80:20, dan 90:10. Sebelum dilakukan proses *training*, pada data dilakukan *sampling* dengan menggunakan SMOTE dalam menangani ketidakseimbangan kelas. Kemudian, dilakukan *training* data menggunakan metode klasifikasi GBT dan diintegrasikan dengan teknik *bagging* guna meningkatkan kinerja model klasifikasi dalam mengatasi *misclassification* dari data tidak seimbang.

1) Gradient Boosted Trees: Gradient Boosting pertama kali diperkenalkan oleh J.H. Friedman [10]. GBT mampu membangun *decision tree* berdasarkan peningkatan dalam struktur pohon pada pembelajaran yang lemah untuk memperbaiki kesalahan pohon dan mencegah terjadinya potensi *overfitting*. Dalam membangun *decision tree*, dapat dilakukan penambahan jumlah iterasi yang sangat konservatif yang dapat menghasilkan dan meningkatkan kinerja model yang lebih baik. GBT mampu memecahkan masalah dengan menyesuaikan pembelajaran lemah dengan gradien negatif dari fungsi kerugian (*loss function*) dan meningkatkan pohon (*trees*) dengan parameter yang mewakili variabel *split* yang dipasang pada setiap *node* terminal pohon.

2) Synthetic Minority Over-sampling Technique: SMOTE merupakan pendekatan tingkatan data yang menggunakan sampel yang berlebih dengan memfokuskan pembelajaran dan bias terhadap distribusi kelas minoritas, sehingga dapat meningkatkan kinerja model klasifikasi. Ilustrasi algoritme SMOTE ditunjukkan pada Gbr. 4 dan proses SMOTE dapat dilakukan pada pembuatan sintetik baru dengan tahapan algoritme sebagai berikut.



Gbr. 4 Ilustrasi algoritme SMOTE.

- Identifikasi fitur vektor dan tetangga terdekatnya.
- Mengambil perbedaan antara kedua hal tersebut.
- Menggandakan perbedaan dengan jumlah acak antara 0 dan 1.
- Identifikasi titik baru pada garis segmen dengan menambahkan jumlah acak ke fitur vektor.
- Mengulangi proses tersebut di atas untuk fitur yang diidentifikasi sampai dengan selesai.

3) Bootstrap Aggregating (Bagging): Teknik *bagging* merupakan algoritme *ensemble* dengan pendekatan sampel *bootstrap* secara acak dengan pembobotan dari kesalahan model klasifikasi (*misclassification*) untuk meningkatkan stabilitas kinerja akurasi dari sebuah model klasifikasi. Penjelasan teknik *bagging* dengan *pseudo-code* algoritme adalah sebagai berikut.

Set data latih $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$,

- Sampel T set elemen n dari D (*replacement*)
  $D_1, D_2, \ldots D_T \rightarrow T$ set pelatihan
- *Training* di setiap D_i, i = 1, …, T dan urutan T *output* $f_1(x)$, …, $f_T(x)$

Aggregate classifier dapat digunakan untuk regresi maupun klasifikasi dengan formula pada (1) sampai (3).
Regresi

$$f(x) = \sum_{i=1}^{T} f_i(x). \quad (1)$$

rata-rata $f_{\text{i untuk } i = 1, \ldots, T}$

Klasifikasi

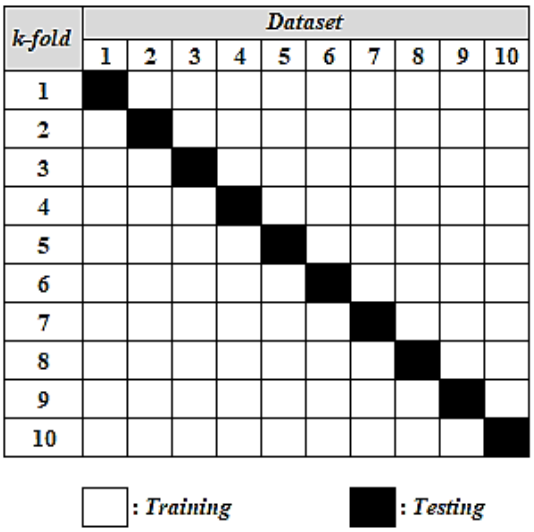
$$f(x) = sign(\sum_{i=1}^{T} f_i(x)) \quad (2)$$

atau

$$f(x) = sign(\sum_{i=1}^{T} sign(f_i(x))). \quad (3)$$

Pada usulan model, validasi dilakukan terhadap model klasifikasi dengan 10 *fold cross-validation*, seperti ditunjukkan pada Gbr. 5, dengan jumlah seluruh data *training* dibagi ke dalam sepuluh bagian. Satu bagian dijadikan sebagai data *testing* dan sembilan lainnya dijadikan sebagai data *training*. Begitu seterusnya secara bergantian dan dari nilai hasil keseluruhan dihitung rata-ratanya [11].

Setelah terbentuk sebuah model, selanjutnya dilakukan pengujian dengan data *testing* sesuai pembagian yang sudah ditentukan. Laporan hasil dari kinerja model diberikan untuk mengetahui kinerja dan perbandingan hasil yang diperoleh.



Gbr. 5 10 *fold cross-validation.*

D. Eksperimen dan Pengujian Model

Pengujian model yang diusulkan dalam eksperimen ini menggunakan spesifikasi *platform* komputer prosesor Intel Core i3-5005U, CPU @2.00 GHz (4 CPUs), 4 GB RAM, dengan sistem operasi Windows 10 64-bit, dan *analytics tools* Weka 3.8 dan Rapidminer 9.3. Eksperimen dan pengujian model ini dilakukan untuk mendapatkan sebuah model dengan pengujian model sebagai *testing* pada model prediksi kelulusan mahasiswa/i.

E. Evaluasi dan Validasi

Pada tahapan ini, dilakukan evaluasi dan validasi. Proses hasil dari sebuah model diukur berdasarkan keakuratan kinerja klasifikasi yang dievaluasi menggunakan matriks kebingungan (*confusion matrix*), seperti yang ditunjukkan pada Tabel II, dengan mengukur nilai *accuracy* dan *Area Under Curve* (AUC) dari sebuah model yang dibangun.

Pada Tabel II disajikan *confusion matrix* yang memiliki prediksi terhadap kelas dengan aktual kelas yang terdiri atas *True Positive* (*TP*) sebagai kelas positif yang telah diidentifikasi dengan benar; *False Positive* (*FP*), yaitu kelas negatif yang telah diidentifikasi salah; *False Negative* (*FN*), yaitu kelas positif yang telah diidentifikasi salah; dan *True Negative* (*TN*) adalah kelas negatif yang telah diidentifikasi benar. Evaluasi dihitung dari hasil *confusion matrix* dengan rumusan seperti pada (1) sampai (6).

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

$$Sensitivity\ (SN)\ = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity\ (SP)\ = \frac{TN}{TN+FP} \quad (6)$$

$$Positive\ Predictive\ Value\ (PPV)\ = \frac{TP}{TP+FP} \quad (7)$$

$$Negative\ Predictive\ Value\ (NPV)\ = \frac{TN}{TN+FN} \quad (8)$$

$$F-measure\ (F)\ = \frac{2TP}{2TP+FP+FN}. \quad (9)$$

TABEL II
CONFUSION MATRIX

Kelas Prediksi	Kelas Aktual	
	Tidak	**Tepat**
Tidak	TP	FP
Tepat	FN	TN

TABEL III
CONFUSION MATRIX GBT

Kelas Prediksi	Kelas Aktual	
	Tidak	**Tepat**
Tidak	340	76
Tepat	44	101

Pada pengujian sebuah model dari metode klasifikasi perlu dilakukan penilaian berdasarkan AUC [12]. Penilaian AUC ditentukan dengan kriteria pernyataan keberhasilan sebagai berikut.

0,90 – 1,00 : sangat baik (*excellent classification*)
0,80 – 0,90 : baik (*good classification*)
0,70 – 0,80 : wajar (*fair classification*)
0,60 – 0,70 : buruk (*poor classification*)
< 0,60 : gagal (*failure*)

III. HASIL DAN PEMBAHASAN

Pada eksperimen ini, pertama-pertama dilakukan pembagian terhadap *dataset* kelulusan mahasiswa dengan jumlah data *training* 70% dan data *testing* 30%, yang akan diterapkan pada metode GBT. Pada eksperimen kedua, dilakukan penanganan masalah pada ketidakseimbangan kelas dengan jumlah data *training* 70% dengan teknik SMOTE dan data *testing* 30% (70:30) yang diterapkan pada metode GBT yang disingkat dengan GBT+SMOTE. Eksperimen ketiga yaitu melanjutkan eksperimen kedua dengan mengintegrasikan dan menerapkan teknik *bagging* yang disingkat dengan GBT+SMOTE+*bagging*.

Eksperimen selanjutnya sama seperti eksperimen pertama, kedua, dan ketiga, tetapi jumlah *dataset* berbeda, yaitu jumlah data *training* 80% dan jumlah data *testing* 20% (80:20). Kemudian, eksperimen berikutnya dengan jumlah data *training* 90% dan jumlah data *testing* 10% (90:10). Hasil eksperimen ditunjukkan pada Tabel III untuk *split* 90:10 dan mewakili contoh proses perhitungan pada eksperimen dengan *split* jumlah data sebelumnya.

Tabel III merupakan *confusion matrix* dengan jumlah data *split* 90:10. Hasil pengujian dari model GBT pada kelas prediksi yang dinyatakan tidak tepat waktu ternyata kelas aktualnya benar tidak tepat waktu, yang berjumlah lebih besar dari hasil prediksi yang lain, dan jumlah yang paling sedikit terdapat pada kelas prediksi tepat waktu yang ternyata kelas aktualnya tidak tepat waktu. Kemudian dilakukan perhitungan nilai kelas prediksi sebagai berikut.

$$Accuracy\ = \frac{340+101}{340+76+44+101} = 78{,}61\ \%$$

$$Sensitivity = \frac{340}{340+44} = 88{,}54\ \%$$

TABEL IV
CONFUSION MATRIX GBT+SMOTE

Kelas Prediksi	Kelas Aktual	
	Tidak	Tepat
Tidak	310	43
Tepat	74	134

TABEL V
CONFUSION MATRIX GBT+SMOTE+*BAGGING*

Kelas Prediksi	Kelas Aktual	
	Tidak	Tepat
Tidak	329	54
Tepat	55	123

$$Specificity = \frac{101}{101+76} = 57{,}06\ \%$$

$$Positive\ Predictive\ Value = \frac{340}{340+76} = 81{,}73\ \%$$

$$Negative\ Predictive\ Value = \frac{101}{101+44} = 69{,}66\ \%$$

$$F-measure = \frac{2x340}{(2x340)+76+44} = 85{,}00\ \%.$$

Pada Tabel IV, hasil *confusion matrix* dari model GBT+SMOTE mengalami perubahan nilai hasil, yakni ada peningkatan pada kelas prediksi tepat waktu dan ternyata kelas aktualnya tepat waktu, dari nilai 101 menjadi 134, sehingga akurasi dapat meningkat dengan perhitungan sebagai berikut.

$$Accuracy = \frac{310+134}{310+43+74+134} = 79{,}14\ \%$$

$$Sensitivity = \frac{310}{310+74} = 80{,}73\ \%$$

$$Specificity = \frac{134}{134+43} = 75{,}71\ \%$$

$$Positive\ Predictive\ Value = \frac{310}{310+43} = 87{,}82\ \%$$

$$Negative\ Predictive\ Value = \frac{134}{134+74} = 64{,}42\ \%$$

$$F-measure = \frac{2x310}{(2x310)+43+74} = 84{,}12\ \%.$$

Pada Tabel V diperlihatkan *confusion matrix* dari GBT+SMOTE+*bagging*. Terdapat peningkatan pada hasil prediksi dari ekperimen sebelumnya pada model GBT+SMOTE, yaitu pada kelas prediksi tidak tepat waktu ternyata kelas aktualnya tidak tepat waktu, dari 310 menjadi 329. Perhitungan terhadap kelas prediksi adalah sebagai berikut.

$$Accuracy = \frac{329+123}{329+54+55+123} = 80{,}57\ \%$$

$$Sensitivity = \frac{329}{329+55} = 85{,}68\ \%$$

$$Specificity = \frac{123}{123+54} = 69{,}49\ \%$$

TABEL VI
RESUME PERBANDINGAN HASIL *CONFUSION MATRIX* 1

Split	**Metode**	***TP***	***FP***	***FN***	***TN***	***ACC***	***AUC***
70:30	GBT	1.021	236	144	282	77,42	0,817
70:30	GBT+SMOTE	919	166	246	352	75,52	0,810
70:30	GBT+SMOTE+*bag*	964	184	201	334	77,12	0,813
80:20	GBT	664	132	116	210	77,90	0,822
80:20	GBT+SMOTE	638	111	142	231	77,45	0,824
80:20	GBT+SMOTE+*bag*	650	105	130	237	79,06	0,824
90:10	GBT	340	76	44	101	78,61	0,848
90:10	GBT+SMOTE	310	43	74	134	79,14	0,849
90:10	GBT+SMOTE+*bag*	329	54	55	123	**80,57**	**0,858**

TABEL VII
RESUME PERBANDINGAN HASIL *CONFUSION MATRIX* 2

Split	**Metode**	***SN***	***SP***	***PPV***	***NPV***	***F***
70:30	GBT	87,64	54,44	81,23	66,20	84,31
70:30	GBT+SMOTE	78,88	67,95	84,70	58,86	81,69
70:30	GBT+SMOTE+*bag*	82,75	64,48	83,97	62,43	83,35
80:20	GBT	85,13	61,40	83,42	64,42	84,26
80:20	GBT+SMOTE	81,79	67,54	85,18	61,93	83,45
80:20	GBT+SMOTE+*bag*	83,33	69,30	86,09	64,58	84,69
90:10	GBT	**88,54**	57,06	81,73	**69,66**	85,00
90:10	GBT+SMOTE	80,73	**75,71**	**87,82**	64,42	84,12
90:10	GBT+SMOTE+*bag*	85,68	69,49	85,90	69,10	**85,79**

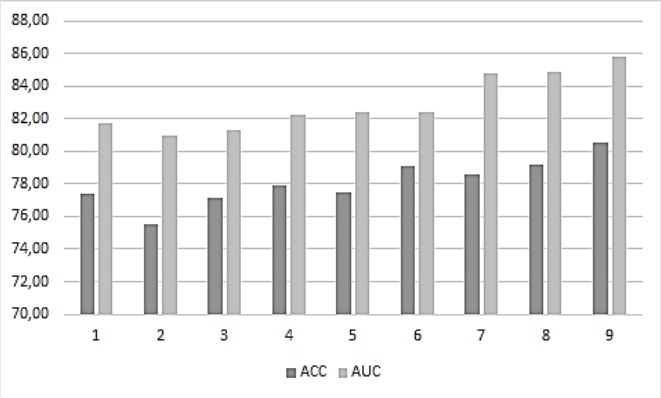
$$Positive\ Predictive\ Value = \frac{329}{329+54} = 85{,}90\ \%$$

$$Negative\ Predictive\ Value = \frac{123}{123+55} = 69{,}10\ \%$$

$$F-measure = \frac{2x329}{(2x329)+54+55} = 85{,}79\ \%.$$

Hasil eksperimen yang dilakukan dari awal hingga akhir dirangkum pada Tabel VI dan Tabel VII sebagai nilai perbandingan. Tabel VI dan Tabel VII merupakan satu kesatuan hasil perhitungan yang dikelompokkan dari tiga kelompok *dataset split* (70:30, 80:20, dan 90:10), sedangkan Gbr. 6 menunjukkan perbandingan nilai *accuracy* dan AUC dalam bentuk grafik. Nilai minimum pada hasil *accuracy* adalah 75,52% dan nilai maksimumnya 80,57%. Untuk AUC, nilai minimumnya 0,810 dan nilai maksimumnya 0,858. Kemudian, nilai rata-rata *accuracy* adalah 78,09% dan nilai rata-rata AUC sebesar 0,829. Pada hasil eksperimen dengan *split* 70:30 menggunakan metode GBT tanpa *improvement*, diperoleh nilai *accuracy* dan AUC yang cukup baik. Dengan *split* 70:30, terjadi penurunan pada model GBT+SMOTE. Sedangkan eksperimen dengan *split* 90:10 untuk *improvement* pada metode GBT+SMOTE+*bagging* menghasilkan peningkatan yang terus menaik, baik pada nilai *accuracy* mapun nilai AUC.

Dari Tabel VI, Tabel VII, dan Gbr. 6 terlihat bahwa terdapat perubahan peningkatan nilai yang cukup signifikan.



Keterangan:

No.	*Split*	Metode	No.	*Split*	Metode
1	70:30	GBT	6	80:20	GBT+SMOTE+*bagging*
2	70:30	GBT+SMOTE	7	90:10	GBT
3	70:30	GBT+SMOTE+*bagging*	8	90:10	GBT+SMOTE
4	80:20	GBT	9	90:10	GBT+SMOTE+*bagging*
5	80:20	GBT+SMOTE			

Gbr. 6 Grafik perbandingan nilai *accuracy* dan AUC (dalam persen).

Peningkatan perolehan nilai secara keseluruhan untuk *accuracy* dan AUC terdapat pada *split* data 90:10, dengan nilai tertinggi *accuracy* sebesar 80,57% dan nilai AUC sebesar 0,858, yaitu pada model GBT+SMOTE+*bagging*. Sementara untuk nilai tertinggi *sensitivity* adalah sebesar 88,54% dan NPV 69,66%, yaitu pada model GBT dengan *split* 90:10. Sedangkan nilai tertinggi *specificity* sebesar 75,71% dan PPV 87,82%, pada model GBT+SMOTE dengan *split* 90:10.

IV. KESIMPULAN

Pada makalah ini, disajikan laporan tentang kinerja model yang dibangun menggunakan metode klasifikasi GBT dengan SMOTE dan *bagging* dalam mengatasi masalah ketidak-seimbangan kelas (*class imbalance*) dan reduksi kesalahan pada model klasfikasi (*misclassification*) dari *dataset* yang tidak seimbang, sehingga meningkatkan kinerja sebuah model GBT. Hasil menunjukkan bahwa nilai terbaik terdapat pada *split* 90:10 dengan nilai *accuracy* sebesar 80,57% dan nilai AUC sebesar 0,858 serta pada *split* 90:10 dengan nilai rata-rata *accuracy* sebesar 79,44% dan nilai *AUC* 0,852.

Dapat disimpulkan bahwa penerapan SMOTE dan *bagging* terbukti mampu memberikan solusi terhadap penanganan masalah ketidakseimbangan kelas dan dapat meningkatkan kinerja model klasifikasi GBT. Berdasarkan *receiver operating characteristic* sebagai AUC, hasil yang diperoleh memiliki kriteria klasifikasi yang baik.

Penelitian selanjutnya tentang deteksi kelulusan mahasiswa/i tepat waktu dan tidak tepat waktu pada masalah *class imbalance* dapat dilakukan pada tingkatan data dengan menggunakan kombinasi teknik *clustering* dan teknik *sampling*.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Direktorat Jenderal Penguatan Riset dan Pengembangan – Kementerian Riset, Teknologi, dan Pendidikan Tinggi yang telah memberikan bantuan hibah dan Universitas Pamulang yang telah memberikan kontribusi pada penelitian.

REFERENSI

[1] A. Bisri dan R.S. Wahono, "Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree," *J. Intell. Syst.*, Vol. 1, No. 1, hal. 27–32, Feb. 2015.

[2] A. Nurhuda dan D. Rosita, "Prediction Student Graduation on Time Using Artificial Neural Network on Data Mining Students STMIK Widya Cipta Dharma Samarinda," *Proc. of the 2017 Int. Conf. on E-commerce, E-Business and E-Government*, hal. 86–89, 2017.

[3] A. Saifudin, "Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang," Jurnal Teknologi, Vol. 10, No. 1, hal. 25–36, 2018.

[4] A. Ali, S.M. Shamsuddin, dan A.L. Ralescu, "Classification with Class Imbalance Problem: A Review," *Int. J. Adv. Soft Comput. Its Appl.*, Vol. 7, No. 3, hal. 176–204, 2015.

[5] N.V. Chawla, D.A. Cieslak, L.O. Hall, dan A. Joshi, "Automatically Countering Imbalance and Its Empirical Relationship to Cost," *Data Min. Knowl. Discov.*, Vol. 17, No. 2, hal. 225–252, 2008.

[6] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, dan F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Trans. Syst. Man, Cybern. Part C (App. and Rev.)*, Vol. 42, No. 4, hal. 463–484, Jul. 2012.

[7] N.V. Chawla, K.W. Bowyer, dan L.O. Hall, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, Vol. 2009, No. Sept. 28, hal. 321–357, 2006.

[8] J.H. Friedman, "Stochastic Gradient Boosting," *Comput. Stat. Data Anal.*, Vol. 38, No. 4, hal. 367–378, 2002.

[9] B.W. Yap, K. Abd-Rani, H.A. Abd-Rahman, S. Fong, Z. Khairudin, dan N.N. Abdullah, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets," *Proc. of the First Int. Conf. on Advanced Data and Information Engineering*, 2014, hal. 13–22.

[10] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, Vol. 29, No. 5, hal. 1189–1232, 2014.

[11] I.H. Witten, E. Frank, M.A. Hall, dan C.J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Ed. San Fransisco, USA: Morgan Kaufmann, 2016.

[12] F. Gorunescu, *Data Mining Concepts, Models and Techniques*, Heidelberg, Germany: Springer-Verlag Berlin Heidelberg, 2011.