Xplore (ISSN:2302-5751), Vol 2 No 2 (2018):32-40

Penanganan Data Tidak Seimbang pada Pemodelan *Rotation Forest* Keberhasilan Studi Mahasiswa Program Magister IPB

Junjun Wijaya*, Agus M Soleh*, Akbar Rizki*
*Departemen Statistika Institut Pertanian Bogor

***Abstrak*—Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB) menyatakan bahwa tidak semua mahasiswa program magister IPB berhasil menyelesaikan studinya. Hal ini menjadi evaluasi untuk IPB agar lebih selektif memilih mahasiswa ke depannya. Penelitian ini bertujuan untuk memodelkan klasifikasi kerberhasilan studi mahasiswa magister IPB tahun 2011 hingga 2015 dengan peubah respon yaitu lulus dan tidak lulus sedangkan profil dan latar belakang mahasiswa sebagai peubah penjelas. Metode klasifikasi yang digunakan yaitu *rotation forest*. Persentase banyaknya mahasiswa yang lulus sangat besar dibandingkan yang tidak lulus, hal ini dapat menyebabkan nilai evaluasi berbeda. SMOTE (*Synthetic Minority Oversampling Technique*) merupakan salah satu metode untuk menangani data tidak seimbang tersebut dengan cara membangkitkan data buatan. Kurva ROC (*Receiver Operating Characteristic*) dibangun untuk melihat nilai *cut off* optimum. Ada dua model klasifikasi, yaitu model *rotation forest* sebelum dan setelah ditangani dengan SMOTE. Hasil perbandingan menunjukkan bahwa model *rotation forest* setelah SMOTE dengan nilai *cut off* 0.6 adalah model terbaik. Model ini mampu meningkatkan nilai sensitivitas lebih dari 50% walaupun nilai akurasi dan spesifisitasnya menurun dibandingkan dengan pemodelan sebelum SMOTE.**

***Kata kunci*—data tidak seimbang; klasifikasi; magister IPB; *rotation forest*; SMOTE**

I. PENDAHULUAN

A. Latar Belakang

Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB) awalnya memiliki tujuh program studi sejak didirikan pada tahun 1975 (IPB (2013)). IPB merupakan kampus yang terus mengembangkan inovasi hingga saat ini SPs-IPB memiliki 9 fakultas dengan 71 program studi magister. SPs-IPB menerima mahasiswa melalui jalur beasiswa dan mandiri. Calon mahasiswa biasanya memilih SPs-IPB dengan beberapa alasan, diantaranya yaitu kompetensi dosen, daya saing lulusan dan standar akademik tinggi. Selain itu, lulusan IPB tersebar di dalam maupun luar negeri, jumlah dan kualitas riset salah satu tertinggi di Indonesia, kerjasama luas, fasilitas lengkap, dan mayor bidang pertanian terlengkap di Asia Tenggara.

Data yang diperoleh dari Basis Data SPs-IPB menyatakan bahwa tidak semua mahasiswa program magister berhasil menyelesaikan studinya. Hal tersebut menjadi evaluasi untuk IPB agar lebih selektif menerima mahasiswa ke depannya. Mahasiswa pascasarjana dinyatakan lulus apabila telah memenuhi persyaratan yang ditetapkan IPB. Syarat tersebut yaitu telah lulus dan menyelesaikan seluruh SKS yang disyaratkan, mendapat nilai minimal B pada penelitian yang sudah diseminarkan, lulus ujian akhir dan menyelesaikan tesis, serta melunasi biaya pendidikan dan administrasi lainnya. Banyak alasan mahasiswa lulus atau tidak lulus dalam melanjutkan studinya. Profil dan latar belakang pendidikan mahasiswa dapat diindikasikan menjadi faktor penyebab keberhasilan studi mahasiswa.

Penelitian ini dilakukan untuk memodelkan klasifikasi data mahasiswa program magister tahun 2011 hingga 2015. Status mahasiswa yaitu lulus dan tidak lulus (*drop out* dan mengundurkan diri) digunakan sebagai peubah respon sedangkan profil dan beberapa latar belakang mahasiswa sebagai peubah penjelasnya. Metode klasifikasi yang digunakan adalah *rotation forest*. Metode ini lebih baik dibandingkan dengan metode pohon gabungan lainnya, seperti *bagging*, *adaboost*, dan *random forest* (Kuncheva and Rodriguez (2007)). Penelitian Raharjo (2016) juga menunjukkan *rotation forest* menghasilkan prediksi klasifikasi yang lebih baik dan konsisten serta efisien dalam waktu komputasi dibandingkan

metode klasifikasi lain yang dibandingkan dalam penelitiannya.

Status mahasiswa yang lulus dan tidak lulus memiliki jumlah yang timpang. Oleh karena itu, data perlu ditangani dengan menggunakan SMOTE *(Synthetic Minority Oversampling Technique)*. Metode ini membangkitkan data buatan untuk kelas minoritas yaitu kategori tidak lulus. Apabila data tidak ditangani, model klasifikasi lebih cenderung memodelkan kelas mayoritas sehingga kontribusi kelas minoritas terhadap model kecil atau sensitivitasnya kecil (Chawla et al. (2002)). Penelitian-penelitian sebelumnya seperti yang dilakukan Fatharani (2016) dan Anindya (2017) terhadap data yang berbeda karakteristiknya menunjukkan bahwa SMOTE dapat meningkatkan akurasi model.

Ada dua model klasifikasi yang diperoleh, yaitu model terhadap data awal dan model setelah ditangani dengan SMOTE. Model klasifikasi terbaik ditentukan dengan mengevaluasi kedua model menggunakan matriks konfusi berdasarkan tingkat akurasi, sensitivitas, dan spesifisitas. Model klasifikasi terbaik diharapkan memiliki nilai akurasi tinggi sehingga dapat digunakan dalam mengambil kebijakan yang tepat.

B. Tujuan

Tujuan dari penelitian ini adalah melakukan pemodelan *rotation forest* terhadap keberhasilan studi mahasiswa program magister IPB tahun 2011 hingga 2015 sebelum dan setelah ditangani dengan SMOTE serta membandingkan kedua model tersebut.

II. TINJAUAN PUSTAKA

A. Rotation Forest

Rotation forest merupakan metode pohon gabungan *(classifier ensembles)* dengan menggunakan analisis komponen utama atau *principal component analysi*s (PCA) untuk merotasi sumbu peubah yang akan dibangun pohon keputusannya. Pohon keputusan digunakan sebagai dasar pengklasifikasian karena sifatnya yang sensitif terhadap rotasi sumbu peubah namun tetap akurat. Meskipun menggunakan analisis komponen utama, semua komponen utama tetap digunakan untuk membangun pohon keputusan agar menjaga kelengkapan informasi data (Rodriguez et al. (2006)).

Misalkan $\underline{\mathbf{x}} = [x_1, ..., x_p]^T$ merupakan vektor amatan dengan p peubah, $\mathbf{X}$ merupakan gugus data dari gabungan vektor $\underline{\mathbf{x}}$ berukuran $n \times p$ dan $\mathbf{F}$ merupakan gugus p peubah. Sedangkan $\underline{\mathbf{y}} = [y_1, ..., y_n]^T$ merupakan vektor kelas dari peubah respon berukuran $n \times 1$. Algortima untuk pembuatan pohon keputusan $D_i; i = 1, 2, ..., U$ adalah sebagai berikut:

1) Bagi $\mathbf{F}$ secara acak menjadi k gugus peubah yang saling lepas dengan banyaknya peubah (m) yang hampir sama. $\mathbf{F}_{i,j}$ merupakan gugus peubah untuk membangun pohon D_i, dengan m_j peubah asal, untuk $j = 1, 2, ..., k$. $\mathbf{X}_{i,j}$ merupakan gugus data $\mathbf{X}$ dengan peubah $\mathbf{F}_{i,j}$.
2) Pada gugus data $\mathbf{X}_{i,j}$ lakukan proses *bootstrap*. Gugus data contoh hasil bootstrap dinyatakan dengan $\mathbf{X}^*_{i,j}$.
3) Lakukan analisis komponen utama pada $\mathbf{X}^*_{i,j}$ dan simpan koefisien komponen utama sebagai $\mathbf{a}^{(1)}_{i,j}, \mathbf{a}^{(2)}_{i,j}, \ldots, \mathbf{a}^{(m_j)}_{i,j}$.
4) Susun vektor-vektor koefisien utama ke dalam sebuah matriks rotasi $\mathbf{R}_i$ seperti dibawah ini:

$$\mathbf{R}_i = \begin{bmatrix} \mathbf{a}^{(1)}_{i,1}, \ldots, \mathbf{a}^{(m_1)}_{i,1} & [\mathbf{0}] & \cdots & [\mathbf{0}] \\ [\mathbf{0}] & \mathbf{a}^{(1)}_{i,2}, \ldots, \mathbf{a}^{(m_2)}_{i,2} & \cdots & [\mathbf{0}] \\ \vdots & \vdots & \ddots & \vdots \\ [\mathbf{0}] & [\mathbf{0}] & \cdots & \mathbf{a}^{(1)}_{i,k}, \ldots, \mathbf{a}^{(m_k)}_{i,k} \end{bmatrix}$$

5) Susun kembali kolom peubah $\mathbf{R}_i$ sehingga bersesuaian dengan susunan gugus peubah. Notasikan matriks rotasi yang telah tersusun kembali dengan $\mathbf{R}^a_i$, berukuran $p \times p$.
6) Bangun pohon keputusan ke-i (D_i) dengan menggunakan $(\mathbf{XR}^a_i, \mathbf{y})$.
7) Ulangi langkah 1) sampai 6) hingga diperoleh U pohon keputusan.

B. SMOTE (Synthetic Minority Oversampling Technique

Chawla et al. (2002) mengusulkan SMOTE sebagai salah satu metode untuk menangani ketidakseimbangan data dengan melakukan *resampling* pada data tersebut. Ide dasar dari SMOTE adalah menambah jumlah contoh pada kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data baru (data sintetis) berdasarkan tetangga terdekat *(k-nearest neighbor)*. SMOTE-NC merupakan SMOTE untuk data yang terdiri dari peubah

numerik dan kategorik (SMOTE-*Nominal Continous*). Penentuan jarak terdekat dihitung dengan jarak Euclidean dengan nilai median dari simpangan baku semua peubah numerik kelas minoritas sebagai selisih nilai peubah kategorik. Formula jarak Euclidean adalah sebagai berikut.

$$\Delta(x, y) = \sqrt{(\vec{x} - \vec{y})^T (\vec{x} - \vec{y})} \quad (1)$$

dengan
$\Delta(x, y)$: jarak antara amatan x dan y
$\vec{x}$: vektor amatan x
$\vec{y}$: vektor amatan y

Tahapan pembangkitan data baru (sintetis) dengan SMOTE setelah menghitung jarak terdekat adalah sebagai berikut:

1) Data Numerik
   Menghitung selisih vektor amatan dari vektor k-tetangga terdekat (*k-nearest neighbor*) dan mengalikannya dengan bilangan acak antara 0 dan 1, kemudian menambahkan hasil tersebut dengan vektor amatan sehingga diperoleh vektor amatan baru. Persamaan untuk menentukan vektor amatan baru dapat dituliskan sebagai berikut.

$$x_{baru} = x + (x^* - x) \times rand[0, 1] \quad (2)$$

   dengan
   x_{baru} : vektor amatan baru
   x : vektor amatan awal
   x^* : vektor amatan k-tetangga terdekat
   $rand[0, 1]$: bilangan acak antara 0 dan 1
2) Data Kategorik
   Menentukan kategori amatan yang paling sering muncul (modus) antara vektor amatan dengan vektor *k-nearest neighbor*, apabila terdapat nilai yang sama maka pilih secara acak. Jadikan nilai yang diperoleh tersebut sebagai amatan baru.

C. Evaluasi Model

Salah satu metode yang dapat digunakan untuk mengukur kinerja suatu sistem klasifikasi adalah matriks konfusi. Matriks konfusi berisi informasi tentang kelas yang sebenarnya dan kelas diprediksi. Pengukuran terhadap kinerja suatu sistem klasifikasi perlu dilakukan agar dapat memberikan gambaran seberapa baik kinerja sistem dalam mengklasifikasikan data dengan benar. Matriks konfusi ditunjukkan pada Tabel I.

Tabel I
MATRIKS KONFUSI

Kelas Prediksi	*Kelas Aktual*	
	Kelas = 0	Kelas = 1
Kelas = 0	A	B
Kelas = 1	C	D

Akurasi dalam klasifikasi adalah persentase ketepatan pelabelan data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Kamber and Han (2001)). Semakin tinggi level akurasinyya, maka dapat dikatakan semakin efektif model algoritma klasifikasi tersebut (Mittal and Gill (2014)). Nilai akurasi, sensitivitas dan spesifisitas dapat dihitung dengan persaman berikut:

$$\text{Akurasi} = \frac{A + D}{A + B + C + D}$$

$$\text{Sensitivitas} = \frac{A}{A + C} \quad \text{Spesifisitas} = \frac{D}{B + D}$$

Kurva *Receiver Operating Characteristic* (ROC) merupakan kurva analisis yang menggambarkan kinerja suatu model klasifikasi pada dua dimensi. *True positive rate* (TPR) pada sumbu-y atau vertikal dan *false positive rate* (FPR) pada sumbu-x atau horizontal. Dalam kurva ROC terdapat garis diagonal yang menghubungkan titik (0,0) dan (1,1). Titik amatan pada kurva ROC yang menunjukkan performa relatif TPR dan FPR diperoleh dengan mengubah nilai *cut off* atau menghitung performa klasifikasi pada beberapa nilai *cut off*. Titik potong atau *cut off* adalah nilai batas antara hasil uji positif dan hasil uji negatif (Fawcett (2006)).

III. METODE PENELITIAN

A. Data

Data yang digunakan adalah data sekunder berupa data mahasiswa program magister Sekolah Pascasarjana Institut Pertanian Bogor angkatan 2011 hingga

2015 yang diperoleh dari Basis Data SPs-IPB. Data yang sesuai kriteria ada sebanyak 4951 amatan dengan 1 peubah respon dan 9 peubah penjelas. Peubah respon (Y) yang digunakan adalah status mahasiswa yang dikategorikan menjadi dua, yaitu lulus dan tidak lulus (*drop out* dan mengundurkan diri). Sedangkan peubah penjelas (X) terdiri dari peubah kategorik yaitu jenis kelamin (X1), status perkawinan (X2), status penerimaan (X3), status perguruan tinggi S1 (X4), sumber biaya pendidikan S2 (X5), kelompok instansi (X6), program studi S2 (X7), dan peubah numerik yaitu usia masuk S2 (X8), dan IPK S1 (X9).

B. Prosedur Analisis Data

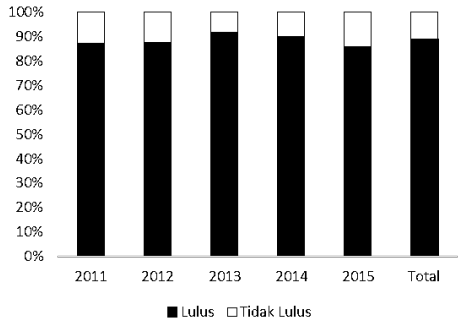
Analisis dalam penelitian ini dibantu dengan *software R Studio Version 1.0.136* dengan bahasa *R Version 3.3.3* menggunakan *package "rotationForest"* dan *"DMwR"*. Tahapan analisis yang dilakukan dalam data penelitian ini adalah sebagai berikut:

1) Melakukan analisis deskriptif untuk mengetahui gambaran umum data mahasiswa program magister
2) Mengubah peubah kategorik menjadi peubah *dummy* karena metode *rotation forest* menggunakan analisis komponen utama, maka setiap peubah harus diubah menjadi peubah numerik
3) Membagi data sebanyak 80% data *training* dan 20% data *testing* secara acak dengan proporsi kategori yang relatif sama dengan data asli
4) Melakukan klasifikasi *rotation forest* untuk memodelkan keberhasilan studi mahasiswa program magister pada data *training* dan melakukan evaluasi kebaikan model pada data *testing* dengan nilai *cut off* standar yaitu 0.5
5) Menangani ketidakseimbangan data dengan SMOTE pada data *training*
    a) Menentukan nilai *k* sama dengan 5 untuk k-tetangga terdekat
    b) Menentukan besar persentase *oversampling* yaitu sebesar 700%
    c) Menghitung jarak antar data kelas minoritas yaitu tidak lulus
    d) Menentukan 5-tetangga terdekat
    e) Melakukan perhitungan untuk membangkitkan data baru (sintetis)
6) Melakukan kembali klasifikasi *rotation forest* untuk memodelkan keberhasilan studi mahasiswa program magister pada data *training* setelah SMOTE dan melakukan evaluasi kebaikan model pada data *testing* dengan nilai *cut off* standar yaitu 0.5
7) Melakukan 100 kali pengulangan pada tahap 3 hingga 6 untuk melihat kestabilan model
8) Membentuk kurva ROC untuk memilih beberapa nilai *cut off* yang sesuai untuk data sebelum dan setelah SMOTE
9) Melakukan kembali tahap 3 hingga 7 namun menggunakan beberapa nilai *cut off* dari kurva ROC pada tahap 8
10) Membandingkan pemodelan klasifikasi *rotation forest* sebelum dan setelah SMOTE pada beberapa nilai *cut off* dengan melihat sebaran nilai akurasi, sensitivitas, dan spesifisitas.

IV. HASIL DAN PEMBAHASAN

A. Deskripsi Data

Gambaran umum mahasiswa program magister Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB) dapat dilihat dengan melakukan eksplorasi data. Banyaknya data mahasiswa yang digunakan adalah 4951 dengan status mahasiswa lulus dan tidak lulus. Gambaran umum status mahasiswa tahun 2011 hingga 2015 dapat dilihat pada Gambar 1 berikut.

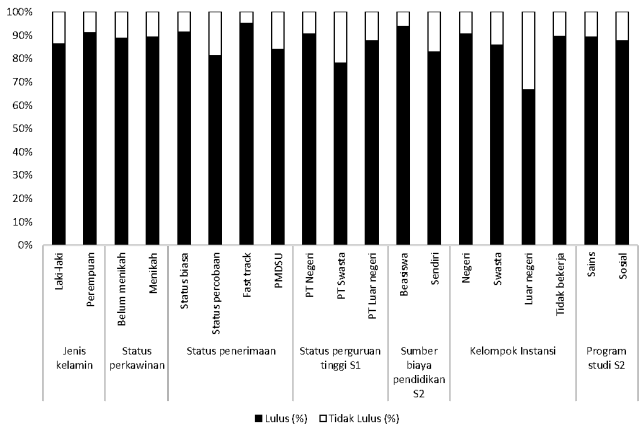


Gambar 1. Persentase status mahasiswa program magister SPs-IPB

Gambar 1 memperlihatkan bahwa persentase total status mahasiswa lulus sebesar 89.03% sedangkan mahasiswa tidak lulus sebesar 10.97%. Status mahasiswa tidak lulus memiliki persentase lebih kecil dibandingkan mahasiswa lulus. Hal inilah yang

menyebabkan data termasuk ke dalam data kelas tidak seimbang. Persentase mahasiswa tidak lulus mengalami penurunan dari tahun 2011 hingga 2013, selanjutnya pada tahun 2013 hingga 2015 mengalami kenaikan. Tahun 2015 merupakan persentase tertinggi mahasiswa tidak lulus sebesar 14.21%, sedangkan persentase terendah pada tahun 2013 sebesar 8.05%.

Persentase status mahasiswa tiap kategori masing-masing peubah penjelas kategorik dapat dilihat pada Gambar 2. Mahasiswa berjenis kelamin perempuan lebih banyak dibandingkan laki-laki, yaitu sebesar 58.07%. Namun mahasiswa berjenis kelamin laki-laki memiliki persentase tidak lulus sebesar 13.82% adalah lebih tinggi dari perempuan. Melihat dari status perkawinan, mahasiswa dengan kategori menikah (89.34%), termasuk duda dan janda memiliki persentase lulus lebih tinggi dibandingkan belum menikah (88.67%). Pada status penerimaan, persentase tertinggi yaitu 18.84% merupakan mahasiswa tidak lulus yang diterima dengan status percobaan sedangkan melalui jalur *fast track* ada pada persentase terendah yaitu sebesar 4.97%.



Gambar 2. Persentase status mahasiswa tiap kategori peubah penjelas

Persentase tidak lulus kategori perguruan tinggi swasta sebesar 21.89%, angka ini terbilang tinggi dilihat dari selisihnya dengan perguruan tinggi luar negeri dan negeri berturut-turut sebesar 9.39% dan 12.35%. Persentase mahasiswa tidak lulus dengan status beasiswa sebesar 6.21%, berbeda jauh dengan mahasiswa yang kuliah dengan biaya sendiri yaitu 17.10%. Berdasarkan pekerjaan mahasiswa, diperoleh bahwa mahasiswa yang bekerja di instansi negeri memiliki persentase tidak lulus lebih rendah. Berbeda mahasiswa yang bekerja di instansi swasta dan tidak bekerja, persentase keduanya adalah 14.24% dan 10.54%. Sedangkan mahasiswa yang bekerja pada instansi luar negeri memiliki persentase tertinggi yaitu 33.33%, namun angka tersebut dikarenakan jumlah mahasiswa dari instansi ini hanya ada 3 orang dan hanya 1 orang yang tidak lulus. Program studi yang mereka jalani memperlihatkan hasil yang berbeda. Persentase mahasiswa tidak lulus program sosial lebih tinggi dibanding sains. Selisih persentase dua program studi tersebut adalah sebesar 1.83%.

Tabel II
DESKRIPSI PEUBAH PENJELAS KATEGORI NUMERIK

Peubah	Rata-rata	Min	Q1	Median	Q3	Maks
Usia	25	19	22	23	27	56
- L	25	19	22	23	26	56
- TL	26	20	23	24	27	56
IPK	3.19	2.00	3.00	3.20	3.48	4.00
- L	3.20	2.00	3.00	3.21	3.48	4.00
- TL	3.13	2.00	3.00	3.13	3.38	3.94

Deskripsi pada Tabel II menunjukkan bahwa usia rata-rata mahasiswa saat masuk program magister adalah 25 tahun, dengan usia paling muda yaitu 19 tahun dan paling tua yaitu 56 tahun. Rata-rata usia mahasiswa lulus yaitu 25 tahun dan tidak lulus yaitu 26 tahun. Indeks Prestasi Kumulatif (IPK) program sarjana mahasiswa berada pada rata-rata 3.19 dengan skala 4. Rata-rata IPK S1 mahasiswa lulus yaitu 3.20 dengan minimum dan maksimum berturut-turut yaitu 2.00 dan 4.00. Sedangkan mahasiswa tidak lulus memiliki rata-rata IPK S1 yaitu 3.13 dengan minimum dan maksimum berturut-turut yaitu 2.00 dan 3.94. Saat menjadi mahasiswa program magister, mereka harus memiliki IPK lebih dari 3.00 dan menyelesaikan studi kurang dari 8 semester sesuai ketentuan yang ditetapkan oleh SPs-IPB.

B. Pemodelan Rotation Forest Sebelum SMOTE

Pemodelan rotation forest dibentuk dari data latih (*training*) sebanyak 80% dan evaluasi model dilakukan pada data uji (*testing*) sebanyak 20% dari total amatan. Komposisi pembagian data *training*

dan *testing* dengan proporsi kelas yang relatif sama dengan data asli. Pemodelan dan evaluasi dilakukan sebanyak 100 kali ulangan. Pada Tabel III diperlihatkan komposisi dari data pada masing-masing status kelulusan mahasiswa.

Tabel III
KOMPOSISI DATA SEBELUM SMOTE

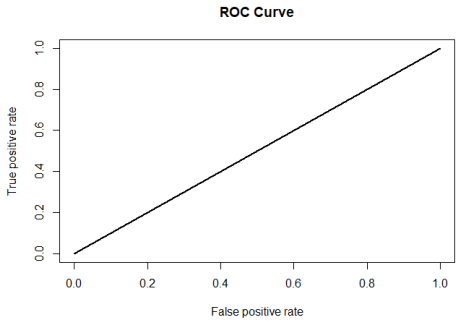
Kinerja klasifikasi	Data *training*	Data *testing*
Tidak lulus	434 (10.96%)	109 (11.00%)
Lulus	3526 (89.04%)	882 (89.00%)
Total	**3960 (100.00%)**	**991 (100.00%)**

Pohon yang dibangun dari pemodelan *rotation forest* ada sebanyak 10 pohon dengan pembagian K *subset* peubah sebanyak 1/3 dari banyaknya peubah penjelas. Hasil ketepatan klasifikasi *rotation forest* dari 100 kali pengulangan dengan nilai *cut off* 0.5 ditunjukkan pada Tabel IV. Nilai sensitivitas sangat kecil artinya semua mahasiswa tidak lulus diprediksi lulus. Sebaliknya, nilai spesifisitas sangat sempurna artinya semua mahasiswa yang lulus benar diprediksi lulus. Perbedaan kedua nilai ini menunjukkan bahwa ketepatan klasifikasi dalam kelas minoritas yaitu tidak lulus sangat rendah untuk memprediksi mahasiswa yang tidak lulus. Hal ini akan menjadi permasalahan dalam mengambil keputusan, sehingga diperlukan penanganan ketidakseimbangan kelas ini untuk memperoleh model *rotation forest* yang lebih baik.

Tabel IV
HASIL KLASIFIKASI ROTATION FOREST SEBELUM SMOTE

Kinerja klasifikasi	Rata-rata (%)
Akurasi	89.00
Sensitivitas	0.00
Spesifisitas	100.00

Kurva ROC pada Gambar 3 menunjukkan nilai sensitivitas (TPR) dan 1-spesifisitas (FPR) pada data sebelum SMOTE. Terlihat bahwa kurva yang dibentuk berupa garis lurus, artinya nilai sensitivitas sama dengan nilai 1-spesifisitas. Hal ini menunjukkan bahwa pemodelan *rotation forest* pada data sebelum SMOTE tidak baik untuk memprediksi. Nilai *cut off* yang optimum adalah 0.5, apabila peluang dari hasil prediksi lebih dari 0.5 maka masuk ke dalam kategori 1 yaitu lulus dan kurang dari 0.5 masuk ke dalam kategori 0 yaitu tidak lulus. Oleh karena itu, tidak perlu dilakukan kembali pemodelan *rotation forest* dengan nilai *cut off* dari kurva ROC karena nilai *cut off* sama.



Gambar 3. Kurva ROC pada data sebelum SMOTE

C. Pemodelan Rotation Forest Setelah SMOTE

Jumlah mahasiswa tidak lulus sebanyak 10.97% dan lulus sebesar 89.03%. Ketidakseimbangan data ini dapat ditangani dengan SMOTE (*Synthetic Minority Oversampling Technique*). Komposisi data setelah melalui tahap SMOTE ditunjukkan pada Tabel V. Kedua kelas menjadi lebih seimbang dibandingkan sebelum ditangani. Pemodelan dilakukan terhadap data tersebut dengan menggunakan banyaknya pohon dan K *subset* dengan nilai *cut off* yang sama seperti pemodelan sebelumnya.

Tabel V
KOMPOSISI DATA SETELAH SMOTE

Kinerja klasifikasi	Data *training*	Data setelah SMOTE
Tidak lulus	434 (10.96%)	3472 (49.61%)
Lulus	3526 (89.04%)	3526 (50.39%)
Total	**3960 (100.00%)**	**7432 (100.00%)**

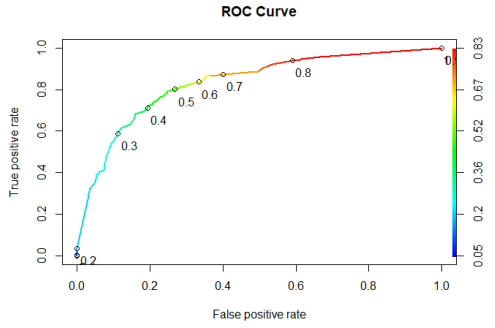
Berdasarkan 100 kali pengulangan, rataan nilai akurasi yang diperoleh pada pemodelan ini sebesar 71.86% dengan rataan nilai sensitivitas dan spesifisitas yaitu 44.40% dan 75.25% seperti yang ditunjukkan pada Tabel VI. Terjadi perubahan yang signifikan, nilai rata-rata akurasi dan spesifisitas menjadi turun sedangkan sensitivitas meningkat dibandingkan pemodelan sebelum SMOTE. Kesalahan prediksi pada kelas minoritas menjadi turun dan tidak cenderung memprediksi kelas mayoritas. Nilai

akurasi, sensitivitas dan spesifisitas hasil pemodelan ini menunjukkan model sudah cukup baik dalam memprediksi mahasiswa yang tidak lulus maupun yang lulus.

Tabel VI
HASIL KLASIFIKASI ROTATION FOREST SETELAH SMOTE

Kinerja klasifikasi	Rata-rata (%)
Akurasi	71.86
Sensitivitas	44.40
Spesifisitas	75.25

Berdasarkan Gambar 4, kurva ROC pada data setelah SMOTE yang dibentuk memiliki daerah yang lebih luas dibanding sebelum SMOTE. Terdapat titik-titik nilai *cut off* dari 0 hingga 1. Maka dipilih beberapa nilai *cut off* yang optimum yaitu 0.3, 0.4, 0.6 dan 0.7. Apabila peluang dari hasil prediksi lebih dari nilai *cut off* maka dikategorikan kelas 1 yaitu lulus sedangkan kurang dari nilai *cut off* maka dikategorikan kelas 0 yaitu tidak lulus. Pemodelan *rotation forest* dilakukan kembali pada masing-masing nilai *cut off* tersebut dan diulang 100 kali.



Gambar 4. Kurva ROC pada data setelah SMOTE

Hasil klasifikasi yang diperoleh ditunjukkan pada Tabel VII. Rata-rata nilai akurasi yang diperoleh menggunakan nilai *cut off* 0.3 adalah 81.84% dengan rata-rata nilai sensitivitas dan spesifisitas yaitu 25.55% dan 88.80%, untuk nilai *cut off* 0.4 memiliki rata-rata nilai akurasi, sensitivitas dan spesifisitas berturut-turut adalah 77.05%, 35.26% dan 82.21%. Sedangkan rata-rata nilai akurasi, sensitivitas dan spesifisitas menggunakan nilai *cut off* 0.6 berturut-turut adalah 67.50%, 51.74% dan 69.45%, dan untuk nilai *cut off* 0.7 memiliki rata-rata nilai akurasi yaitu 57.43% dengan rata-rata nilai sensitivitas yaitu 60.12% dan spesifisitas yaitu 57.10%. Secara umum dapat dilihat bahwa semakin besar nilai *cut off* yang digunakan maka rata-rata nilai akurasi dan spesifisitas menurun dan sebaliknya nilai sensitivitas meningkat.

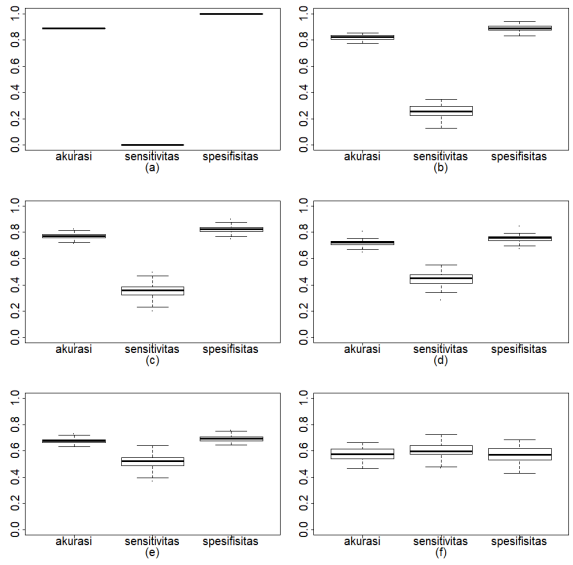
Tabel VII
HASIL KLASIFIKASI ROTATION FOREST SETELAH SMOTE BEBERAPA NILAI CUT OFF

Hasil	*Cut off* 0.3 Rata-rata (%)	*Cut off* 0.4 Rata-rata (%)	*Cut off* 0.6 Rata-rata (%)	*Cut off* 0.7 Rata-rata (%)
Ak	81.84	77.05	67.50	57.65
Sen	25.55	35.26	51.74	60.03
Spe	88.80	82.21	69.45	57.36

D. Perbandingan Pemodelan Rotation Forest

Pemodelan *rotation forest* menghasilkan nilai klasifikasi yang berbeda dari setiap pemodelan yang dilakukan karena data *training* dan *testing* yang digunakan setiap pemodelan dan evaluasi berbeda pula. Oleh karena itu, untuk melihat sebaran nilai klasifikasi maka dilakukan pemodelan dengan pengulangan sebanyak 100 kali. Perbedaan pemodelan sebelum dan setelah SMOTE pada masing-masing nilai *cut off* dapat dilihat dari masing-masing sebaran nilai akurasi, sensitivitas dan spesifisitas yang diperoleh.

Gambar 5 memperlihatkan bahwa nilai akurasi, sensitivitas dan spesifisitas yang diperoleh sebelum SMOTE menghasilkan nilai yang sama dari setiap pengulangan. Sedangkan setelah SMOTE dengan nilai *cut off* yaitu 0.3 hingga 0.7, nilai akurasi, sensitivitas dan spesifisitas sangat menyebar. Artinya, keragaman nilai klasifikasi setelah SMOTE lebih tinggi. Semakin besar nilai *cut off* maka semakin meningkat nilai sensitivitas dan sebaliknya nilai akurasi dan spesifisitas semakin menurun. Hal ini ditunjukkan dari posisi diagram kotak garis diatas dengan bentuk yang hampir simetris untuk semua nilai *cut off*. Nilai median akurasi dan spesifisitas yang dihasilkan dari pemodelan *rotation forest* sebelum SMOTE lebih tinggi dibandingkan setelah SMOTE. Berbeda dengan nilai median sensitivitas sebelum SMOTE yang mengalami peningkatan sesuai nilai *cut off* yang digunakan.



Gambar 5. Diagram kotak garis sebaran nilai akurasi, sensitivitas, dan spesifisitas pemodelan *rotation forest* sebelum SMOTE (a) dan setelah SMOTE pada nilai *cut off* (b) 0.3, (c) 0.4, (d) 0.5, (e) 0.6, dan (f) 0.7

Pemodelan *rotation forest* setelah SMOTE dengan nilai *cut off* 0.6 dapat lebih meningkatkan nilai sensitivitas, walaupun nilai akurasi dan spesifisitas menurun. Data yang dibangkitkan merupakan data sintesis pada kategori tidak lulus sehingga nilai sensitivitas tidak lebih besar dari nilai spesifisitas. Pemodelan *rotation forest* pada data setelah SMOTE dengan nilai *cut off* 0.6 adalah lebih baik dibanding nilai *cut off* lainnya. Hal ini dikarenakan model tidak cenderung memprediksi ke salah satu kategori atau lebih mampu memprediksi mahasiswa tidak lulus.

V. SIMPULAN

Pemodelan *rotation forest* sebelum SMOTE menghasilkan nilai akurasi, sensitivitas dan spesifisitas berturut-turut adalah 89%, 0% dan 100%. Hasil pemodelan ini sangat buruk sehingga perlu dilakukan penanganan. Pada pemodelan ini dilakukan penanganan menggunakan SMOTE. Pemodelan *rotation forest* pada data setelah SMOTE menghasilkan sensitivitas (44.40%) meningkat, namun nilai akurasi (71.86%) dan spesifisitas (75.25%) menurun. Kurva ROC dibangun untuk melihat nilai *cut off* yang sesuai untuk masing-masing data. Nilai *cut off* 0.3, 0.4, 0.6, dan 0.7 dipilih untuk dilakukan pemodelan kembali. Pemodelan *rotation forest* setelah SMOTE dengan nilai *cut off* 0.6 adalah model terbaik dilihat dari nilai akurasi, sensitivitas dan spesifisitas berturut-turut yaitu 67.50%, 51.74% dan 69.45%. Hal ini dikarenakan model tidak cenderung memprediksi ke salah satu kategori atau lebih mampu memprediksi mahasiswa tidak lulus.

DAFTAR PUSTAKA

Anindya, A. (2017). *Penerapan SMOTE pada Metode CART untuk Penanganan Data Tidak Seimbang (Studi Kasus: Klasifikasi Pengangguran dan Bukan Pengangguran di Provinsi Banten [skripsi]*. Bogor(ID): Institut Pertanian Bogor. In press.

Chawla, V., K. Bowyer, and W. Kegelmeyer (2002). Smote: Synthetic minority oversampling technique. *Journal of Atrificial Intelligence Research 9(1)*, 321–357.

Fatharani, R. (2016). *Penerapan Synthetic Minority Oversampling Technique pada Pemodelan Regresi Logistik Ordinal Berat Lahir Bayi di Provinsi Jawa Timur [skripsi]*. Bogor(ID): Institut Pertanian Bogor. In press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters 27*, 861–874.

IPB (2013). *Katalog Program Pascasarjana IPB*. Bogor(ID): Institut Pertanian Bogor.

Kamber, M. and J. Han (2001). *Data mining: Concepts and Techniques*. San Francisco (US): Morgan Kaufmann Publishers.

Kuncheva, L. and J. Rodriguez (2007). An experimental study on rotation forest ensembles. *Lecture Notes in Computer Science 4472*, 459–468.

Mittal, P. and N. Gill (2014). A comparative analysis of classification techniques on medical datasets. *IJRET: International Journal of Research in engineering and Technology 3(6)*, 454–460.

Raharjo, M. (2016). *Kajian Empirik Akurasi Prediksi Klasifikasi Metode Rotation Forest [skripsi]*. Bogor(ID): Institut Pertanian Bogor. In press.

Rodriguez, J., L. Kuncheva, and C. Alonso (2006). Rotation forest: A new classifier ensemble

method. *IEEE Transactions on Pattern Analysis and Machine Intellgince 28(10)*, 1619–1630.