Xplore (2021), Vol. 10, No. 02, 152-166
DOI: 10.29244/xplore.v10i2.238

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Penerapan *Synthetic Minority Oversampling Technique* pada Pemodelan Regresi Logistik Biner terhadap Keberhasilan Studi Mahasiswa Program Magister IPB*

Mega Pradita Pangestika[1], I Made Sumertajaya[2‡], Akbar Rizki[3]

[123]Department of Statistics, IPB University, Indonesia
‡corresponding author: imsjaya@apps.ipb.ac.id

Abstract

The Postgraduate School of IPB has academic standards as well as high competitiveness of graduates who have spread both at home and abroad. In this study Binary Logistic Regression method was used to determine the factors that influence the success of the study of Postgraduate students of Bogor Agricultural University (Graduate School-IPB). The data used are data from IPB Graduate School students who graduated from 2011 to 2015. The response variable used is the success status of student studies namely graduating and not passing and using 9 explanatory variables namely gender, marital status, admission status when entering S2, college status S1 level, the source of S2 education costs, group of agencies working, S2 study program groups, age when entering S2 and S1 GPA. The data obtained is not balanced with the percentage of students who graduate is greater than those who did not pass, so the imbalance of data is handled with SMOTE if it is not handled it will cause a misclassification. Comparison of classification results seen in testing data. The results in the model before SMOTE have an area under the curve or AUC of 0.6760, an accuracy value of 88.77%, a sensitivity value of 99.09% and a specificity of 4.63%. The model after 600% oversampling SMOTE has an AUC value of 0.6642, an accuracy value of 78.36%, a sensitivity value of 83.65%, and a specificity value of 35.18%. Although the accuracy of the model and sensitivity value before SMOTE was higher than the model after SMOTE, the specificity in the model after SMOTE was higher, which meant that the model after SMOTE was better at predicting minority classes (not graduating).

Keywords: binary logistic regression, SMOTE, unbalance data.

1. Pendahuluan

1.1 Latar Belakang

Pendidikan Pascasarjana Institut Pertanian Bogor (SPs-IPB) dimulai pada tahun 1975 dengan 7 jurusan. Menurut IPB (2011) pada tahun-tahun berikutnya tumbuh jurusan-jurusan baru sesuai dengan perkembangan sumber daya yang ada, terutama

* Received: Feb 2019; Reviewed: May 2021; Published: May 2021

bertambahnya tenaga pengajar yang berhasil menempuh studi Pascasarjana di dalam maupun diluar negeri. Alasan penting mahasiswa memilih sekolah pascasarjana IPB diantaranya adalah kualifikasi akademik, kompetensi dosen tinggi, standar akademik tinggi, mempunyai daya saing lulusan yang tinggi dan lulusan tersebar di dalam maupun di luar negeri. Selain itu IPB memiliki jumlah dan kualitas riset salah satu yang tertinggi di Indonesia dan kerjasama yang luas dengan perguruan tinggi dan lembaga internasional. Sebagai upaya untuk mempertahankan reputasi tersebut, salah satunya dengan cara melakukan evaluasi untuk mengetahui potensi keberhasilan studi mahasiswa program magister saat menempuh proses pendidikannya. Hal ini bertujuan agar mahasiswa yang berpotensi mengalami kegagalan studi dapat diberikan suatu tindakan pencegahan terjadinya kegagalan studi masa mendatang oleh pihak terkait.

Data yang terdapat pada SPs-IPB menunjukkan bahwa tidak semua mahasiswa program magister IPB berhasil lulus. Mahasiswa Pascasarjana dinyatakan lulus apabila memenuhi syarat yang ditetapkan IPB. Keberhasilan studi mahasiswa dipengaruhi beberapa faktor seperti jenis kelamin, status menikah, status penerimaan, status perguruan tinggi S1, sumber biaya, kelompok instansi bekerja, kelompok program studi S2, usia masuk S2 dan IPK S1. Profil dan latar belakang mahasiswa diindikasikan menjadi faktor yang dapat mempengaruhi keberhasilan studi.

Penelitian ini dilakukan untuk mengetahui faktor-faktor yang mempengaruhi tingkat keberhasilan studi mahasiswa SPs-IPB. Peubah respon yang digunakan adalah tingkat keberhasilan studi (lulus atau tidak lulus), sehingga analisis Regresi Logistik Biner digunakan pada penelitian ini. Data yang diperoleh merupakan data yang tidak seimbang di mana mayoritas mahasiswa SPs-IPB lulus dan hanya sedikit yang tidak lulus. Hal ini dapat menyebabkan masalah jika tidak ditangani, karena prediksi model yang dihasilkan akan cenderung kepada kelompok mayoritas sehingga kontribusi kelas minoritas terhadap model kecil (Chawla NV (2002)). Oleh karena itu, SMOTE digunakan untuk mengatasi permasalahan tersebut. Selanjutnya, akan dibandingkan model tanpa SMOTE dan setelah SMOTE.

1.2 Tujuan Penelitian

a. Melakukan pemodelan regresi logistik biner terhadap keberhasilan studi mahasiswa program magister IPB sebelum dan setelah ditangani oleh SMOTE dan membandingkan kedua model tersebut.
b. Mengetahui faktor-faktor yang mempengaruhi keberhasilan studi mahasiswa SPs-IPB.

2. Metodologi

2.1 Data

Data yang digunakan adalah data sekunder berupa data mahasiswa program magister Institut Pertanian Bogor angkatan 2011 hingga 2015 yang diperoleh dari SPs-IPB. Data terdiri dari 4951 amatan. Peubah respon yang digunakan adalah status kelulusan mahasiswa program magister dengan banyak mahasiswa lulus (Y=1) dan tidak lulus (Y=0). Mahasiswa yang berstatus tidak lulus mencakup mahasiswa yang mengundurkan diri dan *Drop Out* (DO). Peubah penjelas yang digunakan sebanyak 9 peubah yang terdiri dari 7 peubah kategorik dan 2 peubah numerik. Peubah-peubah tersebut yaitu, jenis kelamin (X1), status menikah (X2), status penerimaan saat masuk S2 (X3), status perguruan tinggi S1 (X4), sumber biaya pendidikan S2 (X5), kelompok

instansi bekerja (X6), kelompok program studi S2 (X7), usia masuk S2 (X8), IPK S1 (X9). Rincian peubah penjelas tersebut dapat dilihat pada Tabel 1.

Tabel 1 Keterangan Peubah yang Digunakan

Peubah	Indikator	Kategori
X1	Jenis Kelamin	0 = Laki-laki
		1 = Perempuan
X2	Status Menikah	0 = Belum Menikah
		1 = Menikah
X3	Status Penerimaan	0 = Status Biasa
		1 = Status Percobaan
		2 = *Fast Track*
		3 = PMDSU
X4	Status Perguruan Tinggi S1	0 = PT Negeri
		1 = PT Swasta
		2 = PT Luar Negeri
X5	Sumber Biaya Pendidikan S2	0 = Beasiswa
		1 = Sendiri
X6	Kelompok Instansi	0 = Negeri
		1 = Swasta
		2 = Luar Negeri
		3 = Tidak Bekerja
X7	Kelompok Program Studi S2	0 = Sains
		1 = Sosial
X8	Usia Masuk S2	Numerik
X9	IPK S1	Numerik

2.2 Prosedur Analisis Data

Tahapan analisis data sebagai berikut:

1) Melakukan eksplorasi data.
2) Membagi data menjadi data *training* dan data *testing* dengan perbandingan 80% dan 20% secara acak.
3) Analisis regresi logistik biner.
   a. Model regresi logistik merupakan model matematika yang dapat digunakan untuk mengetahui hubungan antar respon dengan satu atau lebih peubah penjelasnya. Apabila peubah respon memiliki dua buah nilai, yaitu sukses dan gagal maka pemodelan statistika yang digunakan adalah regresi logistik biner (Agresti (2007)). Peubah respon memiliki dua kategori yaitu mahasiswa lulus yang diberikan kode 1 dan tidak lulus dengan kode 0 dengan 9 peubah penjelas yang dapat dilihat pada Tabel I. Model regresi logistik biner dengan p peubah penjelas adalah sebagai berikut (Hosmer DW (2000)):

$$\pi(x) = \frac{exp\ exp\ g(x)}{1 + exp\ exp\ g(x)} \quad (1)$$

dimana

$$\pi(x) = P(X = x) = 1 - P(Y = 0|X = x) \quad (2)$$

dan $0 \leq \pi(x) \leq 1$

Setara dengan transformasi logitnya sebagai berikut:

$$logit[\pi(x)] = g(x) = ln\frac{\pi(x)}{1+\pi(x)} \quad (3)$$

$$g(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (4)$$

dengan

p = banyaknya peubah penjelas

$X_1, X_2, \ldots, X_p$ = peubah penjelas

$\beta_0, \beta_1, \ldots, \beta_p$ = parameter yang tidak diketahui dan perlu diduga nilainya

b. Mencari nilai dugaan parameternya. Pendugaan parameter regresi logistik dengan menggunakan metode kemungkinan maksimum. Fungsi kemungkinan maksimumnya adalah:

$$L(\beta) = \prod_{i=1}^{K} [\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}] \quad (5)$$

dengan

$i = 1,2,\ldots,n$

y_i= respon pada amatan ke-i

$\pi(x_i)$= peluang kejadian amatan ke-i bernilai y_i

Secara matematika rumus diatas lebih mudah dikerjakan dalam bentuk log yang ditulis sebagai berikut:

$$L(\beta) = \sum_{i=1}^{k} y_i \, ln \, ln \, [\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)] \quad (6)$$

c. Menguji parameter secara simultan. Pengujian terhadap parameter pada model dilakukan sebagai upaya untuk memeriksa peranan peubah penjelas yang ada di dalam model. Menurut Hosmer DW (2000), untuk mengetahui peran seluruh peubah penjelas di dalam model secara simultan dapat digunakan statistik uji-G.

Hipotesis yang diuji adalah:

$H_0 = \beta 1 = \beta 2 = \ldots = \beta P = 0$

H_1 = paling sedikit ada satu $\beta i \neq 0, i = 1,2,\ldots,p$

Statistik uji-G didefinisikan sebagai:

$$G = -2ln\frac{L0}{LP} \quad (7)$$

dimana L0 adalah fungsi kemungkinan maksimum tanpa peubah penjelas, dan LP merupakan fungsi kemungkinan maksimum dengan p peubah penjelas. Hipotesis nol ditolak jika nilai $G > \chi^2$ (Hosmer DW 2000).

d. Menguji parameter secara parsial. Uji parameter secara parsial yang digunakan adalah statistik uji Wald, dengan hipotesis yang diuji:

$H_0 = \beta i = 0$

$H_1 = \beta i \neq 0, i = 1,2,\ldots,p$

Statistik uji Wald didefinisikan sebagai:

$$W = \frac{B_i}{\widehat{SE}_{B_i}} \quad (8)$$

dengan

$i = 1,2,\ldots,p$

B_i= dugaan koefisien parameter ke-i

$\widehat{SE}_{B_i}$= dugaan galat baku dari B_i

Hipotesis nol ditolak jika nilai $Wald > Z_{a/2}$ (Hosmer DW (2000)).

4) Mengevaluasi model dengan melihat nilai AUC dan tabel kesesuaian klasifikasi pada data *testing* dengan tetap mempertimbangkan kesesuaian klasifikasi pada data *training*. *Area Under Curve* (AUC) adalah luas dibawah kurva yang dalam hal ini merupakan kurva *Receiver Operating Character* (ROC). Menurut Fawcett (2006), kurva ROC menggambarkan performa pengklasifikasian secara dua dimensi. Kurva tersebut adalah plot peluang salah *negative* (1-spesifisitas) pada sumbu X dengan prediksi benar positif (sensitivitas) pada sumbu Y. Jika ingin membandingkan beberapa performa pengklasifikasian maka ROC dapat diubah dalam bentuk skalar salah satunya menjadi AUC. AUC adalah suatu bagian dari daerah satuan persegi yang nilainya antara 0 hingga 1. Kekuatan nilai-nilai model prediksi untuk membedakan antara kasus positif dan negatif diukur dengan area dibawah kurva ROC (Hosmer DW (2000)).

Ukuran kebaikan suatu metode klasifikasi dapat dievaluasi dari tabel kesesuaian klasifikasi yang diterapkan pada data prediksi dan data sebenarnya. Benar negatif dan benar positif merupakan frekuensi amatan yang diprediksi dengan tepat. Salah negatif adalah frekuensi amatan yang sesungguhnya positif diprediksi negative, sedangkan salah positif adalah frekuensi amatan yang sesungguhnya negative diprediksi positif. Evaluasi hasil klasifikasi dapat dilakukan dengan menghitung nilai akurasi, spesifisitas dan sensitivitas. Akurasi menggambarkan tingkat ketepatan klasifikasi secara keseluruhan. Spesifisitas menggambarkan akurasi pada kelas negatif, sedangkan sensitivitas menggambarkan akurasi pada kelas positif. Rincian formula dari ketiga nilai tersebut dapat dilihat pada Tabel 2.

Tabel 2 Kesesuaian Klasifikasi

Prediksi		Hasil Prediksi: Negatif	Hasil Prediksi: Positif	Ketepatan
Hasil Sebenarnya	Negatif	True Negative (TN)	False Positive (FN)	Spesifisitas=TN/(TN+FP)
	Positif	False Negative (FN)	True Positive (TP)	Sensitivitas=TP/(TP+FN)
Akurasi				(TP+TN)/TN+TP+FN+FP)

5) Melakukan SMOTE pada data tidak seimbang.

Synthetic Minority Oversampling Technique (SMOTE) adalah metode yang diusulkan atau diperkenalkan oleh Chawla NV (2002). 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*). Jumlah k-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya. Pembangkitan

data buatan yang berskala numerik berbeda dengan kategorik. Data numerik diukur jarak kedekatannya dengan jarak Euclidean sedangkan data kategorik dengan rumus *Value Difference Metric* (VDM).

Tahapan melakukan SMOTE sebagai berikut:

a. Menghitung jarak antar amatan pada kelas minor menggunakan rumus VDM yaitu (Cost S (1993)):

$$\Delta(X,Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r \quad (9)$$

dengan

$\Delta(x,y)$: jarak antara amatan x dengan y

$W_x W_y$: bobot amatan (dapat diabaikan)

N : banyak peubah penjelas

R : bernilai 1 (jarak Manhattan) atau 2 (jarak Euclidean)

$\delta(x_i, y_i)^r$: jarak antar kategori dengan rumus:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \quad (10)$$

dengan

$\delta(V_1, V_2)$: jarak antara nilai V_1 dan V_2

C_{1i} : banyaknya V_1 yang termasuk kelas i

C_{2i} : banyaknya V_2 yang termasuk kelas i

i : banyaknya kelas i = 1,2,...,m

C_1 : banyaknya nilai 1 terjadi

C_2 : banyaknya nilai 2 terjadi

N : banyaknya kategori

K : konstanta (biasanya 1)

b. Menentukan nilai K=5 dan persentase *oversampling* sebesar 600%.

c. Dipilih satu contoh dari kelas minor secara acak.

d. Menentukan amatan k tetangga terdekat dengan mengurut jarak contoh terpilih dengan semua amatan pada kelas minor.

e. Melakukan perhitungan untuk membangkitkan data baru (sintetis) dengan menentukan nilai per peubah penjelasnya. Nilai tersebut diperoleh dari mayoritas nilai pada k tetangga terdekat. Jika semua peubah telah dibuat maka diperoleh satu amatan baru.

f. Langkah c hingga e dilakukan hingga banyaknya *oversampling* yang diinginkan telah tercapai.

6) Membangun model dengan data yang telah melalui tahap SMOTE dan menguji parameternya.

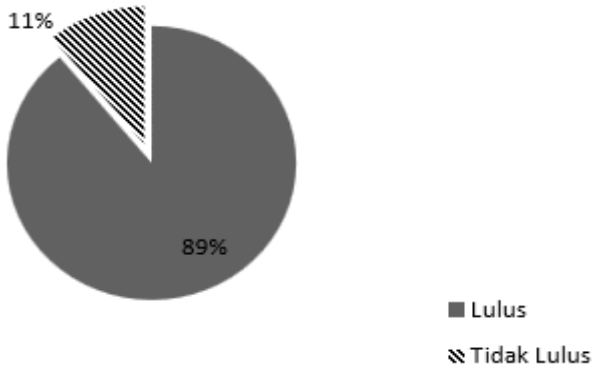
7) Membandingkan hasil model yang dihasilkan tanpa SMOTE dan setelah SMOTE dengan melihat nilai AUC dan tabel kesesuaian klasifikasi data *testing* dengan tetap mempertimbangkan dari data *training* masing-masing model.

3. Hasil dan Pembahasan

3.1 Eksplorasi Data

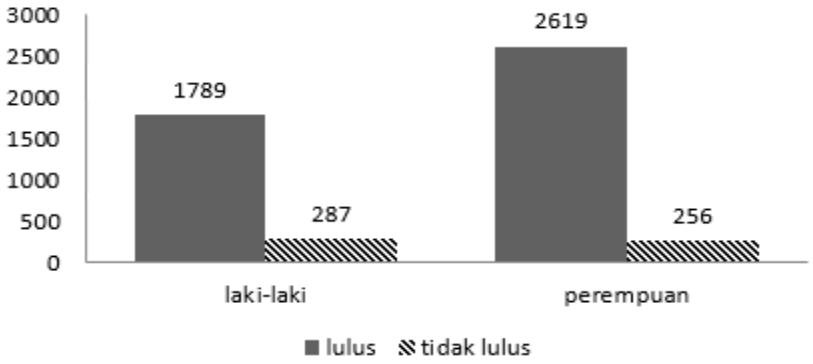
Status kelulusan mahasiswa program magister dapat dilihat pada Gambar 1 yang menunjukkan bahwa persentase mahasiswa yang berstatus lulus sebesar 4408 (89%)

dan mahasiswa yang berstatus tidak lulus sebesar 543 (11%). Jumlah mahasiswa yang berstatus lulus lebih banyak dibanding yang berstatus tidak lulus. Hal ini mengindikasikan bahwa adanya ketidakseimbangan kelas respon pada data mahasiswa program magister IPB.

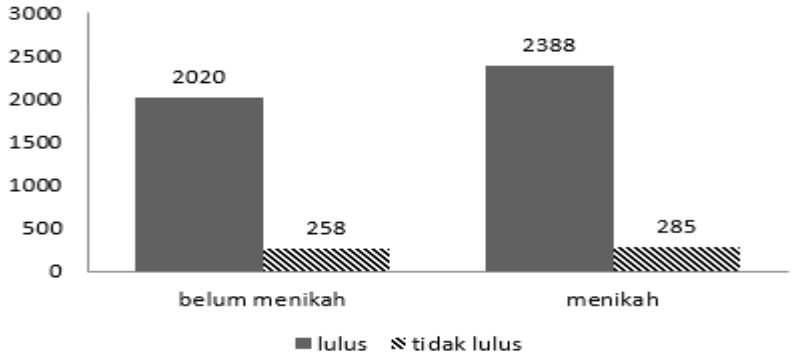


Gambar 1. Diagram lingkaran status kelulusan mahasiswa magister IPB

Gambar 2 menunjukkan status kelulusan berdasarkan jenis kelamin. Informasi yang diperoleh dari gambar 3 menunjukkan bahwa pada peubah jenis kelamin, mahasiswa yang tidak lulus lebih besar terjadi pada kategori laki-laki sebesar 287 (5.79%) mahasiswa dibandingkan kategori perempuan sebesar 256 (5.17%) mahasiswa. Hal ini juga menunjukkan bahwa pada peubah jenis kelamin laki-laki dengan perempuan tidak terdapat perbandingan yang cukup jauh.



Gambar 2. Status kelulusan berdasarkan jenis kelamin

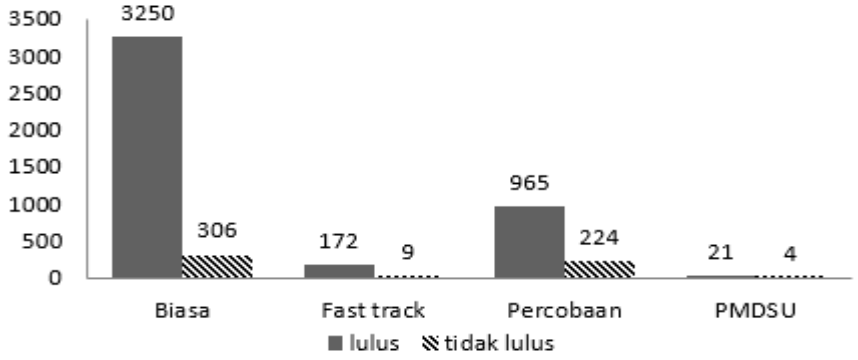


Gambar 3. Status kelulusan berdasarkan status menikah

Informasi yang diperoleh dari Gambar 3 menunjukkan bahwa pada peubah status menikah, mahasiswa yang belum menikah dengan yang sudah menikah memiliki

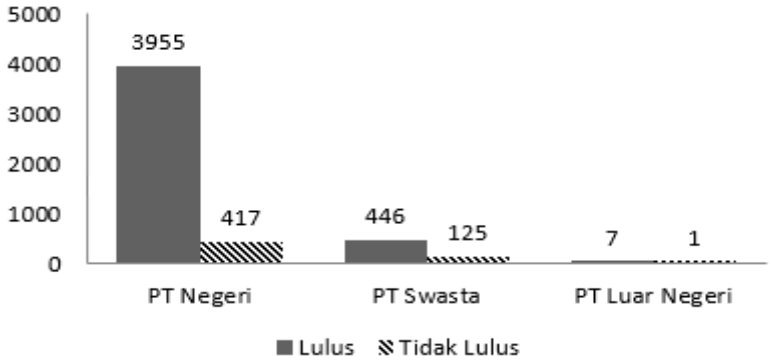
perbandingan jumlah mahasiswa tidak lulus yang tidak terlalu jauh begitu pula dengan perbandingan jumlah mahasiswa yang lulusnya. Jumlah yang tidak lulus pada mahasiswa yang belum menikah sebesar 258 (5.21%) mahasiswa sedangkan yang sudah menikah memiliki jumlah yang tidak lulus sebesar 285 (5.76%) mahasiswa. Hal ini menyatakan bahwa mahasiswa yang sudah menikah memiliki jumlah tidak lulus yang lebih tinggi dibandingkan yang belum menikah.

Status kelulusan berdasarkan status penerimaan ditunjukkan oleh Gambar 4. Mahasiswa dengan jalur biasa memiliki jumlah yang tidak lulus paling tinggi dibanding dengan jalur penerimaan lainnya yaitu *fast track*, percobaan dan PMDSU dengan jumlah sebesar 306 (6.18%) mahasiswa diikuti dengan jalur percobaan sebesar 224 (4.52%) mahasiswa, jalur fast track sebesar 9 (0.18%) mahasiswa dan jalur PMDSU dengan nilai sebesar 4 (0.08%) mahasiswa. Hal ini menunjukkan jumlah tidak lulus tertinggi pada status penerimaan biasa dan yang terendah adalah pada jalur penerimaan PMDSU.



Gambar 4. Status kelulusan berdasarkan status penerimaan

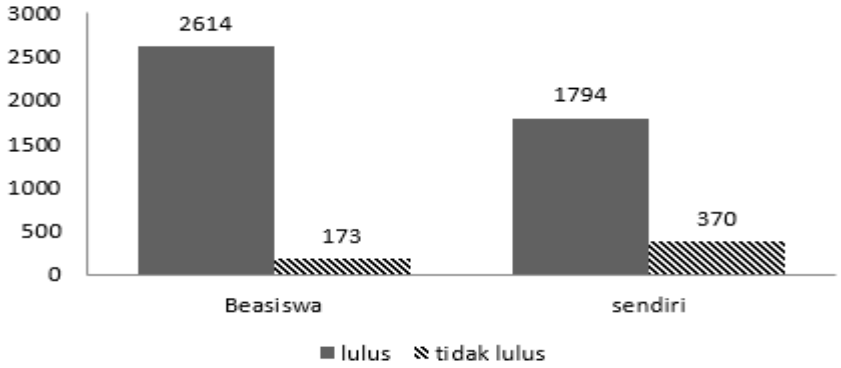
Gambar 5 menunjukkan status kelulusan berdasarkan status perguruan tinggi S1. Informasi yang diperoleh dari gambar 6 adalah jumlah mahasiswa yang tidak lulus tertinggi sampai terendah berturut-turut terjadi pada kategori perguruan tinggi negeri dengan jumlah sebesar 417 (9.51%) mahasiswa, perguruan tinggi swasta dengan jumlah sebesar 125 (2.52%) mahasiswa dan perguruan tinggi luar negeri yang yang berjumlah 1 (0.02%) mahasiswa. Hal ini menunjukkan terdapat perbedaan yang sangat jauh pada jumlah mahasiswa yang tidak lulus antara perguruan tinggi luar negeri dengan perguruan tinggi swasta dan negeri.



Gambar 5. Status kelulusan berdasarkan status perguruan tinggi S1

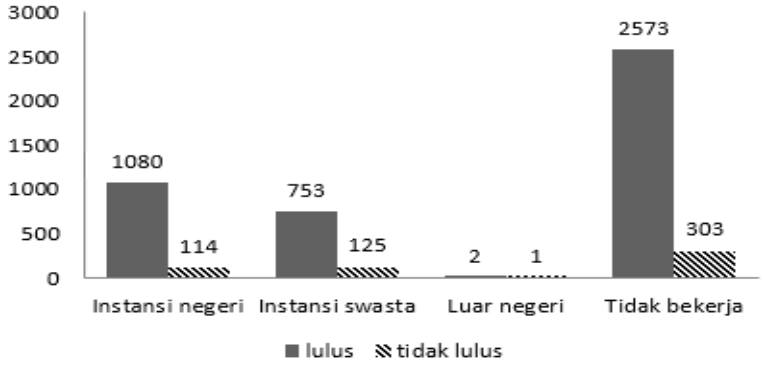
Status kelulusan berdasarkan sumber biaya ditunjukkan oleh Gambar 6. Gambar 6 menunjukkan bahwa jumlah mahasiswa yang tidak lulus pada kategori biaya pendidikan pribadi lebih tinggi jumlahnya dengan nilai sebesar 370 (7.47%)

dibandingkan mahasiswa yang menerima beasiswa yang memiliki jumlah sebesar 173 (3.49%). Perbandingan jumlah mahasiswa tidak lulus berdasarkan biaya pribadi dengan mahasiswa yang menerima beasiswa jumlahnya tidak begitu jauh begitupun dengan perbandingan jumlah pada mahasiswa yang lulus.

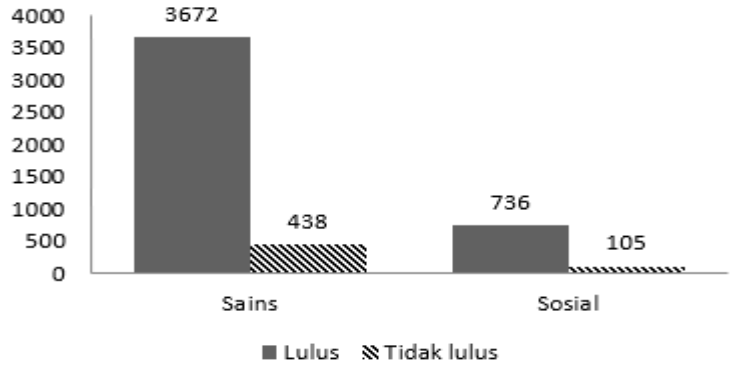


Gambar 6. Status kelulusan berdasarkan sumber biaya

Informasi yang ditunjukkan oleh Gambar 7 menunjukkan bahwa jumlah mahasiswa yang tidak lulus tertinggi yaitu pada mahasiswa yang tidak bekerja sebesar 303 (6.12%) dibanding kategori lainnya yaitu instansi swasta, negeri dan luar negeri dengan jumlah mahasiswa berturut-turut sebesar 125 (2.52%), 114 (2.30%), 1 (0.02%). Instansi luar negeri memiliki jumlah mahasiswa yang tidak lulus paling kecil. Mahasiswa yang tidak lulus pada kelompok instansi bekerja luar negeri memiliki jumlah yang berbeda jauh dengan instansi lainnya.



Gambar 7. Status kelulusan berdasarkan instansi bekerja



Gambar 8. Status kelulusan berdasarkan kelompok program studi

Gambar 8 menunjukkan status kelulusan berdasarkan kelompok program studi S2.

Informasi yang diperoleh adalah mahasiswa yang kuliah pada kelompok sains memiliki jumlah yang tidak lulus lebih besar dibanding mahasiswa yang kuliah di program studi sosial dengan perbandingan jumlah sebesar 438 (8.85%) pada program studi sains dan 105 (2.12%) pada program studi sosial. Secara umum dapat dilihat bahwa persentase mahasiswa yang tidak lulus pada setiap peubah penjelas kategorik memiliki persentase yang jauh lebih kecil dibandingkan dengan mahasiswa yang lulus.

Peubah penjelas numerik yang digunakan dalam penelitian ini adalah usia masuk S2 dan Indeks Prestasi Kumulatif (IPK) S1. Deskripsi dapat dilihat pada Tabel 3, mahasiswa yang berstatus lulus dan tidak lulus masing-masing memiliki rata-rata umur sekitar 25 dan 26 tahun saat diterima program magister dengan rentang umur sekitar 19 sampai 56 tahun dan 20 sampai 56 tahun. Mahasiswa yang berstatus lulus pada program magister IPB memiliki rata-rata IPK 3.21 dengan rentang 2.00 sampai 4.00 sedangkan pada mahasiswa yang berstatus tidak lulus memiliki rata-rata IPK 3.14 dengan rentang 2.00 sampai 3.94.

Tabel 3 Deskripsi Peubah Penjelas Numerik

Stat mhs	Peubah	Min	Q1	Med Kanonik	Rata”	Q3	Maks
Lulus	Usia	19	22	23	25	26	56
	IPK S1	2.00	3.00	3.21	3.21	3.5	4
Tidak lulus	Usia	20	23	24	26	27	56
	IPK S1	2.00	3.00	3.13	3.14	3.4	3.94

3.2 Analisis Regresi Logistik Biner

Analisis Regresi Logistik Biner pada penelitian ini menggunakan 80% data *training* sebesar 3962 amatan sedangkan 20% pada data *testing* sebesar 989 amatan yang digunakan untuk klasifikasi model. Data *training* dan *testing* tersebut memiliki persentase yang sama untuk kategori mayor sebesar 89% dan untuk kategori minor sebesar 11%. Dugaan parameter yang dihasilkan dari model yang dibangun pada data *training* dengan memaksimumkan nilai fungsi kemungkinan secara bersama dengan uji G. Nilai uji G menunjukkan hasil sebesar 244.16 dengan nilai-P bernilai 0.00 < 0.05 yang artinya minimal ada satu peubah penjelas yang berpengaruh terhadap keberhasilan studi mahasiswa program magister IPB. Hasil pendugaan parameter, uji Wald, dan nilai-P dapat dilihat pada Tabel 4. Pengujian parameter secara parsial menggunakan uji Wald menunjukkan bahwa peubah penjelas yang berpengaruh terhadap keberhasilan studi mahasiswa program magister IPB pada taraf nyata 5% adalah jenis kelamin, status penerimaan, dan sumber biaya sedangkan peubah yang lainnya tidak berpengaruh karena nilai-P lebih besar dari 0.05.

Evaluasi model dilakukan dengan membuat tabel kesesuaian klasifikasi. Tabel kesesuaian klasifikasi model sebelum SMOTE pada data *training* dan *testing* dapat dilihat pada Tabel 5. Hasil klasifikasi pada data *training* menghasilkan nilai luas dibawah kurva atau AUC sebesar 0.72, nilai akurasi sebesar 88.59%, nilai sensitivitas sebesar 98.69% dan nilai spesifisitasnya sebesar 6.67%. Hasil klasifikasi pada data *testing* memiliki nilai AUC sebesar 0.67, nilai akurasi sebesar 88.77%, nilai sensitivitas sebesar 99.09% serta nilai spesifisitasnya sebesar 4.63%. Hasil tersebut memperlihatkan bahwa nilai AUC yang diperoleh pada model tanpa SMOTE untuk

data *training* sebesar 0.72 dan data *testing* sebesar 0.67 yang artinya bahwa nilai AUC tersebut sudah mengklasifikasikan model dengan baik. Akan tetapi, nilai spesifisitasnya berbeda sangat jauh dibandingkan dengan nilai sensitivitasnya. Hal ini menunjukkan bahwa klasifikasi cenderung memprediksi kategori mayor dan mengabaikan kategori minor sehingga terdapat kesalahan ketepatan klasifikasi pada kategori minor. Oleh karena itu, kategori minor atau kelas yang tidak lulus perlu diperhatikan karena dapat dikatakan model sebelum SMOTE belum mampu mengklasifikasikan mahasiswa yang tidak lulus dengan baik, sehingga memerlukan penanganan terhadap kategori data yang tidak seimbang agar hasilnya menjadi lebih baik.

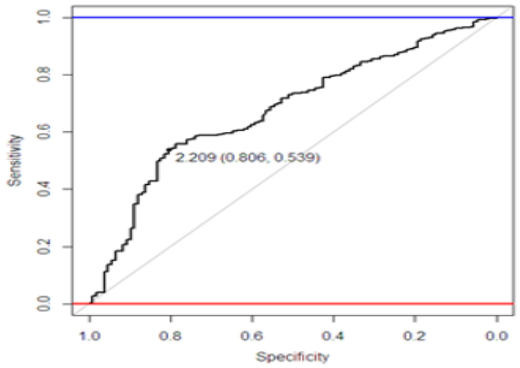
Tabel 4 Dugaan Parameter Model tanpa SMOTE

Peubah Penjelas	B	SE	Wald	p-value
Intersep	15.09	382.97	0.04	0.97
Jenis Kelamin (*X*1)				
perempuan	0.38	0.1	3.52	0.00*
Status Menikah (*X*2)				
Menikah	0.12	0.1	1.15	0.25
Status Penerimaan (*X*3)				
Percobaan	-0.67	0.11	-5.86	0.00*
Fast Track	0.23	0.4	0.58	0.56
PMDSU	-1.66	0.57	-2.92	0.00*
Kelompok PT S1 (*X*4)				
PT Negeri	-12.18	382.97	-0.03	0.97
PT Swasta	-12.97	382.97	-0.03	0.97
Sumber Biaya (*X*5)				
Sendiri	-1.1	0.12	-9.42	0.00*
Kelompok Instansi (X6)				
Swasta	-0.07	0.17	-0.45	0.65
Luar Negeri	-0.58	962.24	-0.001	0.99
Tidak Bekerja	-0.1	0.16	-0.65	0.52
Program Studi (X7)				
Sosial	-0.13	0.14	-0.98	0.32
Umur Masuk S2 (*X*8)	0	0.01	-0.5	0.62
IPK S1 (*X*9)	0.04	0.14	0.26	0.79

Tabel 5 Hasil Klasisfikasi Model tanpa SMOTE

		Prediksi					
		training			*testing*		
		0	1	Ktptn	0	1	Ktptn
Sebenarnya	0	29	406	6.67%	5	103	4.63%
	1	46	3481	98.69%	8	873	99.09%
Akurasi				88.59%			88.77%

Kurva ROC pada Gambar 9 memiliki nilai AUC sebesar 0.67 dan nilai sensitivitas sebesar 99.09%. Nilai sensitivitas ini menunjukkan bahwa metode yang digunakan mampu memprediksi mahasiswa yang lulus (kelas mayoritas) dengan ketepatan yang cukup tinggi yaitu mencapai 99.09%. Akan tetapi, nilai spesifisitasnya tergolong kecil yaitu sebesar 4.63%. Hal ini memperlihatkan bahwa metode belum mampu memprediksi mahasiswa yang tidak lulus (kelas minoritas) dengan baik.



Gambar 9. Kurva ROC model tanpa SMOTE

3.3 Regresi Logistik Biner Setelah SMOTE

Tahapan awal SMOTE adalah melakukan proses pembangkitan data buatan (*resampling*). Beberapa *oversampling* telah dilakukan dari 100% hingga 725%. Hasil *oversampling* yang dipilih yaitu 600% pada data *testing* karena dilihat dari tujuan penelitian agar model tidak cenderung memprediksi kategori mayor yaitu dengan melihat nilai spesifisitas yang lebih tinggi. Sedangkan pada data *training* nilai spesifisitas lebih tinggi terdapat pada *oversampling* 700%. Hasil klasifikasi yang dipilih yaitu dengan melihat data *testing*. Hasil klasifikasi pada model setelah SMOTE *oversampling* 600% pada data *testing* memiliki nilai AUC sebesar 0.66, nilai akurasi sebesar 78.36%, nilai sensitivitas sebesar 83.65%, dan nilai spesifisitas sebesar 35.18%. Hasil klasifikasi pada model setelah SMOTE *oversampling* 600% pada data *training* memiliki nilai AUC sebesar 0.77, nilai akurasi sebesar 75.58%, nilai sensitivitas sebesar 84.78% serta nilai spesifisitasnya sebesar 52.44%. Berdasarkan hasil klasifikasi tersebut bahwa nilai AUC pada model setelah SMOTE *oversampling* 600% pada data *testing* menurun menjadi lebih rendah yang artinya AUC pada model setelah SMOTE sudah lebih baik karena berdasarkan pada pertimbangan dengan melihat nilai spesifisitas, sensitivitas dan akurasi totalnya. Berdasarkan nilai spesifisitasnya, nilainya meningkat yang artinya model setelah SMOTE sudah mampu mengklasifikasikan mahasiswa yang tidak lulus dengan baik sesuai tujuan dari penelitian. Data yang telah melalui tahap SMOTE sebanyak 12227 dengan persentase

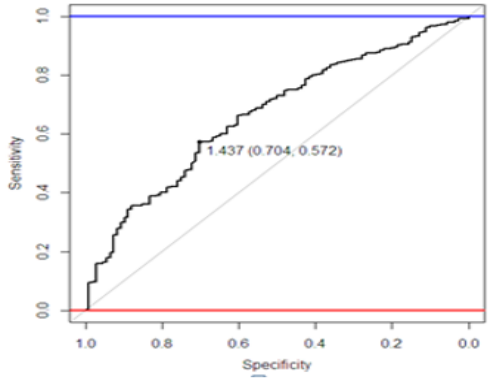
yang tidak lulus sebesar 37% dan yang lulus sebesar 63%. Kelas minoritas pada data yang telah melalui tahap SMOTE menjadi meningkat persentasenya. Setelah itu, data yang telah melalui tahap SMOTE ini dibangun dengan model regresi logistik biner. Pengujian parameter menggunakan uji G menghasilkan nilai sebesar 2459.5 dengan nilai-P bernilai 0.00 < 0.05. Hal ini menyatakan bahwa pada pengujian parameter secara simultan pada taraf nyata 5% untuk model dengan *oversampling* 600% menunjukkan bahwa minimal ada satu peubah penjelas yang berpengaruh terhadap keberhasilan studi mahasiswa program magister IPB. Selanjutnya, dilakukan pengujian secara parsial dengan melihat nilai dari uji Wald. Peubah penjelas yang memiliki nilai-P kurang dari 0.05 maka peubah penjelas tersebut berpengaruh terhadap keberhasilan studi mahasiswa program magister IPB.

Tabel kesesuaian klasifikasi model setelah SMOTE dapat dilihat pada Tabel 6. Ketepatan klasifikasi yang dihasilkan antara data *training* dan data *testing* tidak jauh berbeda, baik pada tingkat akurasi total, spesifisitas maupun sensitivitas. Selain itu nilai sensitivitas dan spesifisitas yang dihasilkan cenderung seimbang, sehingga dapat dikatakan bahwa klasifikasi dari model setelah SMOTE sudah cukup baik.

Tabel 6 Klasifikasi setelah SMOTE *oversampling* 600%

		Prediksi					
		training			*testing*		
		0	1	Ktptn	0	1	Ktptn
Sebenarnya	0	1825	1655	52.44%	38	70	35.18%
	1	1331	7416	84.78%	144	737	83.65%
Akurasi				75.58%			78.36%

Kurva ROC pada Gambar 10 menunjukkan kurva ROC setelah SMOTE *oversampling* 600%. Kurva ROC pada Gambar 10 memiliki nilai AUC sebesar 0.66. Nilai spesifisitas pada model setelah SMOTE meningkat menjadi 35.18% dibandingkan dengan model tanpa SMOTE sebesar 4.63%. Nilai sensitivitas pada model setelah SMOTE sebesar 83.65%. Jika dibandingkan dengan model tanpa SMOTE nilai sensitivitas setelah SMOTE lebih rendah. Akan tetapi nilai spesifisitasnya meningkat sehingga metode yang digunakan telah mampu memprediksi mahasiswa yang tidak lulus (kelas minoritas) lebih baik dibandingkan tanpa SMOTE.



Gambar 10. Kurva ROC model setelah SMOTE *oversampling* 600%

3.4 Perbandingan Model

Perbandingan model dilihat dari model tanpa SMOTE dan model setelah SMOTE dengan *oversampling* 600%. Kedua model tersebut dibandingkan dengan melihat nilai

ketepatan klasifikasi dari tabel kesesuaian klasifikasi yaitu dengan melihat nilai spesifisitas, sensitivitas, akurasi serta nilai AUC. Perbandingan ini dilakukan pada data *testing*. Nilai akurasi pada model tanpa SMOTE sebesar 88.77% sedangkan pada model setelah SMOTE sebesar 78.36% yang artinya model tanpa SMOTE lebih besar nilai akurasinya dibanding model setelah SMOTE. Nilai spesifisitas atau ketepatan model dalam mengklasifikasikan mahasiswa yang tidak lulus pada model setelah SMOTE lebih tinggi dibandingkan dengan model tanpa SMOTE. Besarnya spesifisitas model tanpa SMOTE sebesar 4.63% dan model dengan SMOTE sebesar 35.18%. Nilai sensitivitas atau ketepatan model dalam mengklasifikasikan mahasiswa yang lulus pada model tanpa SMOTE sebesar 99.09% dan pada model setelah SMOTE sebesar 88.42%. Nilai AUC pada model setelah SMOTE lebih rendah 0.01 dibandingkan dengan nilai AUC tanpa SMOTE.

Tabel 7 Perbandingan Evaluasi Model

Evaluasi	Model	
	Tanpa SMOTE	Over 600%
Spesifisitas	4.62%	35.18%
Sensitivitas	99.09%	83.65%
Akurasi	88.77%	78.36%
AUC	0.676	0.6642

3.5 Model Klasifikasi Tingkat Kelulusan

Model klasifikasi yang digunakan adalah model setelah SMOTE pada persentase *oversampling* 600% karena memiliki nilai spesifisitas yang lebih baik sesuai dengan tujuan penelitian. Model dibangun dari tujuh peubah penjelas yang berpengaruh terhadap keberhasilan studi mahasiswa program magister IPB yaitu jenis kelamin, status menikah, status penerimaan saat diterima S2, sumber biaya, kelompok instansi bekerja, kelompok program studi S2 dan IPK S1. Model logit yang diperoleh sebagai berikut:

$$\begin{aligned}\hat{g} = {} & 13.46 + 0.44X_1(Perempuan) + 0.12X_2(Menikah) - 0.86X_3(Percobaan) \\ & + 0.13X_3(Fastrack) - 1.05X_3(PMDSU) - 0.72X_5(Sendiri) \\ & - 0.04X_6(Swasta) - 0.79X_6(Luar\ Negeri) + 0.27X_6(Tidak\ Bekerja) \\ & - 0.88X_7(Sosial) + 0.15X_9(IPK\ S1)\end{aligned}$$

Rasio odds digunakan untuk membandingkan kejadian sukses dan tidak sukses antar kategori peubah bebas. Pada penelitian ini, mahasiswa yang lulus merupakan kejadian sukses sedangkan mahasiswa yang tidak lulus (mengundurkan diri atau *Drop Out*) merupakan kejadian gagal. Interpretasi nilai rasio odds dilakukan pada peubah penjelas yang berpengaruh secara signifikan. Nilai rasio odds dapat dilihat pada Lampiran 1. Nilai rasio odds untuk peubah penjelas jenis kelamin perempuan memiliki peluang lulus 1.55 kali dibandingkan jenis kelamin laki-laki. Nilai rasio odds untuk peubah status menikah memiliki peluang lulus 1.13 kali dibandingkan dengan yang belum menikah. Nilai rasio odds untuk peubah status penerimaan percobaan pada saat diterima S2 memiliki peluang lulus 2.38 kali dibandingkan penerimaan biasa. Nilai rasio odds untuk peubah sumber biaya dengan biaya sendiri memiliki peluang lulus 2.04 kali dibandingkan dengan yang menerima beasiswa. Nilai rasio odds untuk

peubah kelompok instansi yang tidak bekerja memiliki peluang lulus 1.30 kali dibandingkan kelompok instansi bekerja negeri. Nilai rasio odds untuk peubah kelompok program studi sosial memiliki peluang lulus 2.44 kali dibandingkan kelompok program studi sains. Nilai rasio odds untuk peubah IPK S1 dengan nilai rata-rata IPK sebesar 3.21 memiliki peluang lulus 1.16 kali dibandingkan dengan nilai rata-rata IPK sebesar 3.14.

4. Simpulan

Nilai akurasi model tanpa SMOTE sebesar 88.77% sedangkan nilai akurasi model setelah SMOTE sebesar 78.36%. Nilai akurasi tersebut mengalami penurunan sebesar 10.41%. Nilai sensitivitas untuk mengklasifikasikan mahasiswa yang lulus pada model setelah SMOTE mengalami penurunan sebesar 15.44%. Namun, nilai spesifisitas untuk mengklasifikasi mahasiswa yang tidak lulus pada model setelah SMOTE lebih tinggi dibandingkan pada model tanpa SMOTE. Besarnya nilai spesifisitas model tanpa SMOTE sebesar 4.63% dan model setelah SMOTE sebesar 35.18%. Hal ini menunjukkan bahwa nilai spesifisitas tersebut mengalami kenaikan sebesar 30.55% yang artinya walaupun nilai akurasi total dan sensitivitasnya menurun tetapi nilai spesifisitasnya meningkat sehingga model telah mampu dalam memprediksi kelas minoritas atau mahasiswa yang tidak lulus. Nilai AUC pada model tanpa SMOTE sebesar 0.67 sedangkan pada model setelah SMOTE sebesar 0.66. Nilai AUC tersebut mengalami penurunan sebesar 0.01 yang berarti nilai AUC setelah SMOTE tersebut lebih akurat dibandingkan model tanpa SMOTE karena dilihat berdasarkan nilai AUC yang paling kecil. Secara keseluruhan evaluasi model tersebut menunjukkan bahwa SMOTE dapat digunakan untuk mengatasi kesalahan klasifikasi pada data tidak seimbang.

Faktor-faktor yang mempengaruhi keberhasilan studi mahasiswa program magister IPB berasal dari model setelah SMOTE dengan persentase *oversampling* 600% yaitu jenis kelamin, status menikah, status penerimaan saat diterima S2, sumber biaya, kelompok instansi bekerja, kelompok program studi S2 dan IPK S1.

Daftar Pustaka

Agresti, A. (2007). *An Introduction to Categorical Data Analysis Second Edition*. New Jersey (US): John Wiley and Sons Inc.

Chawla NV, Bowyer KW, H. L. K. W. (2002). Smote: Synthetic minority oversampling technique. *Journal of Atrificial Intelligence Research 9(1)*, 321–357.

Cost S, S. S. (1993). A weighted neighbour algorithm for learning with symbolic features. *Machine Learning 10*, 57–58.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letter 2*, 861–874.

Hosmer DW, L. S. (2000). *Applied Logistic Regression : Second Edition*. New York (US): John Wiley and Sons Inc.

IPB (2011). *Katalog Program Pascasarjana IPB*. Bogor (ID): Institut Pertanian Bogor.