Xplore (ISSN:2302-5751), Vol 2 No 1 (2018):33-40

Pemodelan *Support Vector Machine* Data Tidak Seimbang Keberhasilan Studi Mahasiswa Magister IPB

Octavia Dwi Amelia*, Agus M Soleh*, Septian Rahardiantoro*
*Departemen Statistika Institut Pertanian Bogor

***Abstrak*—Salah satu cara yang dapat dilakukan Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB) untuk mempertahankan reputasinya adalah dengan menerapkan sistem penerimaan mahasiswa baru yang lebih selektif. Penelitian ini memprediksi keberhasilan studi mahasiswa menggunakan pemodelan *Support Vector Machine* (SVM) dengan mempertimbangkan karakteristik dan latar belakang pendidikan mahasiswa. Namun pada data yang dimiliki terdapat ketidakseimbangan kelas data. Pemodelan SVM pada data tidak seimbang menghasilkan kinerja yang kurang baik dengan nilai sensitivitas 0.00%. Penanganan data tidak seimbang menggunakan *Sythetic Minority Oversampling Technique* (SMOTE) berhasil meningkatkan kinerja SVM dalam mengklasifikasikan mahasiswa yang tidak lulus. Jenis SVM yang tepat digunakan untuk melakukan pemodelan keberhasilan studi mahasiswa berdasarkan nilai akurasi, sensitivitas, dan spesifisitas dengan *cut off default* adalah SVM RBF. Ketika menggunakan nilai *cut off* terbaik dari setiap jenis SVM, nilai sensitivitas berhasil untuk ditingkatkan kembali. SVM RBF masih tetap memberi hasil yang paling baik ketika menggunakan *cut off* 0.6. Model akhir yang digunakan untuk memprediksi keberhasilan studi mahasiswa SPs-IPB diperoleh dari pemodelan SVM RBF dengan *cut off* 0.6 menggunakan keseluruhan data yang telah melalui tahap SMOTE.**

***Kata kunci*—SMOTE; SPs-IPB; SVM;**

I. Pendahuluan

A. Latar Belakang

SPs-IPB dapat menerapkan sistem penerimaan mahasiswa baru dengan lebih selektif untuk mempertahankan reputasinya. Hal ini dimaksudkan agar mahasiswa yang diterima adalah mahasiswa yang unggul dan berpotensi untuk lulus dari SPs-IPB, sehingga dapat mengurangi jumlah mahasiswa yang tidak lulus. Oleh karena itu, perlu dilakukan pemodelan untuk mendeteksi mahasiswa yang tidak lulus. Kajian mengenai keberhasilan studi mahasiswa, dalam hal ini lulus tidaknya mahasiswa dari SPs-IPB sebelumnya pernah dilakukan oleh Permatasari (2009) menggunakan regresi logistik biner. Sedangkan pada penelitian ini analisis dilakukan menggunakan pemodelan klasifikasi *Support Vector Machine* (SVM) dengan mempertimbangkan karakteristik dan latar belakang pendidikan mahasiswa. SVM merupakan salah satu algoritme yang paling berpengaruh dan banyak digunakan dalam *data mining*. Menurut Wang and Japkowicz (2010), metode ini memiliki dasar teoritis yang sangat kuat. Selain itu, SVM juga memiliki kemampuan generalisasi yang sangat baik (Wu et al. (2008)).

Berbeda dengan penelitian sebelumnya, penelitian ini memperhatikan adanya aspek ketidakseimbangan data. Data dikatakan tidak seimbang karena mahasiswa yang tidak lulus jumlahnya jauh lebih sedikit dibandingkan mahasiswa yang lulus. Kasus seperti ini dapat mengakibatkan salah klasifikasi pada kelas minoritas, yaitu kelas dengan jumlah amatan yang jauh lebih sedikit (Bunkhumpornpat et al. (2012)). Sehingga mahasiswa yang seharusnya tidak lulus akan diprediksi lulus. Hal ini tentunya akan merugikan SPs-IPB jika menerima mahasiswa yang sebenarnya tidak lulus. Oleh karena itu penanganan pada data tidak seimbang perlu dilakukan. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan salah satu metode penanganan yang dapat digunakan. SMOTE dilakukan dengan menambahkan data buatan pada kelas minoritas sehingga data menjadi seimbang. Penelitian yang dilakukan oleh Agwil (2015) menunjukkan bahwa penerapan SMOTE dapat meningkatkan kinerja klasifikasi SVM dalam mengklasifikasikan kelas minoritas.

B. Tujuan

Tujuan dari penelitian ini adalah menerapkan klasifikasi SVM pada data keberhasilan studi mahasiswa program magister IPB dalam memprediksi mahasiswa yang tidak lulus tanpa dan dengan melalui tahap SMOTE.

II. Tinjauan Pustaka

A. Ketidakseimbangan Kelas Data

Data dikatakan tidak seimbang ketika suatu kelas data memiliki jumlah amatan yang jauh lebih sedikit dibandingkan kelas lainnya (Bunkhumpornpat et al. (2012)). Kelas dengan amatan yang sedikit disebut kelas minoritas, sedangkan kelas lainnya disebut kelas mayoritas. Analisis klasifikasi secara umum kurang memadai dalam mengatasi ketidakseimbangan kelas data karena algoritme dibuat tanpa memperhatikan hal tersebut (Han et al. (2005)). Akibatnya kelas minoritas akan mengalami salah klasifikasi. Oleh karena itu perlu dilakukan penanganan. Penanganan dapat dilakukan pada tingkat data dengan mengubah sebaran data yang tidak seimbang menggunakan metode *resampling*. Terdapat dua metode utama dalam melakukan *resampling*, yaitu *undersampling* dan *oversampling*.

B. Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode *oversampling* kelas minoritas dengan menciptakan data buatan (sintetik) untuk mengatasi masalah ketidakseimbangan kelas data (Chawla et al. (2002)). *Oversampling* dilakukan dengan memanfaatkan konsep *k*-tetangga terdekat. Ketika semua data bertipe numerik, jarak tetangga terdekat dihitung menggunakan jarak Euclid. Ketika data bertipe numerik dan kategorik, perhitungan jarak tetap menggunakan jarak Euclid namun menggunakan nilai median dari simpangan baku peubah numerik sebagai selisih nilai peubah kategorik. Nilai median ini dihitung ketika nilai kategori amatan dan tetangga terdekatnya berbeda. Berikut ini merupakan prosedur pembangkitan data buatan.

1) Data Numerik
   Hitung jarak antara vektor amatan dengan vektor *k*-tetangga terdekat dan kalikan jarak tersebut dengan bilangan acak antara 0 sampai 1. Lalu tambahkan hasil perkalian tersebut dengan vektor amatan sehingga diperoleh vektor amatan baru.
2) Data Kategorik
   Amatan baru merupakan kategori yang paling sering muncul pada vektor amatan dan *k*-tetangga terdekatnya. Jika nilainya sama maka dipilih secara acak.

C. Support Vector Machine

Support Vector Machine (SVM) merupakan analisis klasifikasi yang menggunakan bidang pemisah (*hyperplane*) dalam melakukan klasifikasi. Pada ruang berdimensi p, bidang pemisah akan berdimensi $p-1$. SVM dapat menghasilkan berbagai kemungkinan bidang pemisah. Bidang pemisah terbaik yaitu yang memiliki *margin* paling besar (*maximal margin hyperplane*). *Margin* adalah jarak terdekat amatan data latih dengan bidang pemisah. Amatan yang memiliki jarak terdekat dengan bidang pemisah disebut *support vector* (James et al. (2013)).

1) SVM Linear: SVM linier diterapkan pada data yang dapat dipisahkan secara linier. Menurut Scolkopf and Smola (2000) fungsi keputusan SVM linier dapat dituliskan sebagai berikut

$$f(x) = sgn(\langle w, x\rangle + b). \tag{1}$$

Bidang pemisah terbaik ditentukan dengan meminimumkan

$$\frac{1}{2}||w||^2 \tag{2}$$

dengan konstrain

$$y_i(\langle w, x_i\rangle + b) \geq 1, \tag{3}$$

w adalah vektor yang tegak lurus dengan bidang pemisah, b adalah jarak antara titik pusat dengan bidang pemisah, $y_i \in \{-1, 1\}$ merupakan label kelas, sedangkan x_i adalah data latih dengan $i = 1, \ldots, m$. Persoalan di atas dapat diselesaikan dengan menggunakan pengali *Langrange* berikut

$$L(x, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i(y_i(\langle x_i, w\rangle + b) - 1), \alpha_i \geq 0 \tag{4}$$

dengan α_i adalah konstanta *lagrange*. Data latih yang terletak pada *margin* dengan nilai $\alpha > 0$ merupakan *support vector*. Fungsi keputusan akhir

hanya dipengaruhi *support vector* dan dapat dituliskan sebagai berikut

$$f(x) = sgn(\sum_{i=1}^{ns} \alpha_i y_i \langle x, x_i \rangle + b) \quad (5)$$

keterangan :
x = data uji
x_i = *support vector*, $i = 1, 2, \ldots, ns$
ns = banyak data yang merupakan *support vector*.

Soft margin classifier digunakan ketika data tidak dapat dipisahkan menggunakan bidang pemisah secara sempurna. Peubah *slack* (ξ) ditambahkan ke dalam fungsi sehingga bidang pemisah terbaik diperoleh dengan meminimumkan

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \quad (6)$$

dengan konstrain

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (7)$$

C merupakan parameter bernilai non-negatif yang menunjukkan penalti akibat pelanggaran saat klasifikasi

2) SVM Non-linier: Data yang tidak dapat dipisahkan secara linier dapat diatasi dengan memperbesar ruang fitur menggunakan fungsi Kernel $K(x, x_i)$(James et al. (2013)). Fungsi keputusan pada persamaan (5) berubah menjadi

$$f(x) = sgn(\sum_{i=1}^{ns} \alpha_i y_i K(x, x_i) + b). \quad (8)$$

Berikut ini merupakan beberapa fungsi Kernel :

1) Polinomial dengan derajat d

$$K(x, x_i) = (1 + \sum_{j=1}^{p} x_j x_{ij})^d \quad (9)$$

2) *Radial Basis Function* (RBF)

$$K(x, x_i) = exp(-\gamma \sum_{j=1}^{p} (x_j - x_{ij})^2) \quad (10)$$

dengan p adalah banyak peubah penjelas, d adalah derajat polinomial berupa bilangan bulat positif dan γ adalah suatu konstanta positif.

D. Ketepatan Klasifikasi

Evaluasi terhadap kinerja klasifikasi dapat dilakukan menggunakan matriks konfusi seperti yang ditunjukkan pada Tabel 1. Ketika terdapat ketidakseimbangan kelas data, kelas minoritas diberi label positif dan kelas mayoritas diberi label negatif (Bekkar et al. (2013)). Berdasarkan matriks konfusi dapat diperoleh nilai akurasi, sensitivitas, dan spesifisitas. Akurasi menghitung proporsi amatan yang diprediksi dengan tepat terhadap seluruh amatan. Sensitivitas mengukur akurasi klasifikasi amatan kelas positif sedangkan spesifisitas mengukur akurasi klasifikasi amatan kelas negatif. Selain ketiga ukuran keakuratan di atas, kurva ROC juga digunakan untuk memperoleh nilai *cut off* yang dapat meningkatkan kinerja klasifikasi.

Tabel I
MATRIKS KONFUSI

Kelas Aktual	Kelas Prediksi: Positif	Kelas Prediksi: Negatif
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

III. METODOLOGI

A. Data

Data yang digunakan dalam penelitian ini merupakan data mahasiswa tahun ajaran 2011/2012 sampai 2015/2016 pada semua program studi magister yang ada di SPs-IPB. Data tersebut diperoleh dari basis data SPs-IPB yang terdiri dari 4951 amatan dengan 543 mahasiswa berstatus tidak lulus dan 4408 mahasiswa berstatus lulus. Peubah respon yang diamati adalah keberhasilan studi mahasiswa dengan kategori tidak lulus dan lulus. Sedangkan peubah penjelasnya ditunjukkan pada Tabel 2.

B. Metode

Penelitian ini dianalisis menggunakan *Microsoft Excel* dan R 3.4.3 dengan paket e1071, DMwR, caret, dan ROCR. Langkah-langkah dalam melakukan analisis adalah sebagai berikut.

1) Melakukan eksplorasi untuk melihat karakteristik data.
2) Membagi data menjadi data latih dan data uji.

Tabel II
DAFTAR PEUBAH PENJELAS

Peubah Penjelas	Keterangan	Tipe Peubah
X1	Jenis Kelamin	Kategorik
X2	Status perkawinan	Kategorik
X3	Status Penerimaan	Kategorik
X4	Status Perguruan Tinggi Asal	Kategorik
X5	Sumber Biaya Pendidikan	Kategorik
X6	Jenis Pekerjaan	Kategorik
X7	Program Studi	Kategorik
X8	Usia Masuk SPs-IPB	Numerik
X9	IPK Asal S1	Numerik

3) Melakukan klasifikasi SVM pada data latih menggunakan SVM linier dan non-linier serta mengevaluasi kinerja klasifikasi pada data uji.
4) Menerapkan SMOTE pada data latih.
   a) Menghitung jarak antar amatan kelas minoritas.
   b) Menentukan 5 tetangga terdekat
   c) Pilih secara acak 1 tetangga terdekatnya.
   d) Melakukan pembangkitkan data buatan.
   e) Mengulangi langkah b sampai d hingga persentase *oversampling* 700% tercapai.
5) Melakukan klasifikasi SVM pada data latih baru yang telah melalui tahap SMOTE serta mengevaluasi kinerja klasifikasi pada data uji.
6) Mengulangi langkah 2 sampai 5 hingga 100 kali.
7) Menentukan nilai *cut off* untuk setiap jenis SVM menggunakan kurva ROC dengan melakukan klasifikasi SVM pada data keseluruhan.
8) Mengulangi langkah 4 dan 5 hingga 100 kali. Klasifikasi menggunakan nilai *cut off* yang diperoleh pada langkah 7.
9) Menentukan *cut off* terbaik untuk setiap jenis SVM berdasarkan hasil kinerja klasifikasi.
10) Membandingkan kinerja klasifikasi SVM sebelum dan sesudah dilakukan penanganan dengan SMOTE serta setelah dilakukan SMOTE menggunakan *cut off* terbaik.

IV. HASIL DAN PEMBAHASAN

A. Gambaran Umum Data

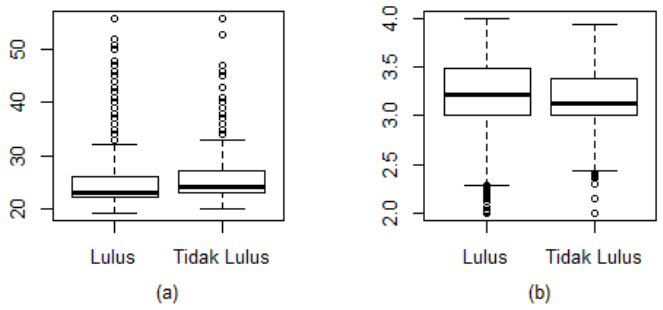
Keberhasilan studi mahasiswa dianalisis menggunakan status kelulusan sebagai peubah responnya. Persentase mahasiswa yang lulus yaitu sebesar 89.03%. Nilai ini jauh lebih besar dibandingkan dengan persentase mahasiswa yang tidak lulus yaitu sebesar 10.97%. Hal ini menunjukkan adanya ketidakseimbangan data. Pemeriksaan ada tidaknya ketidakseimbangan pada data yang akan dianalisis perlu dilakukan karena jika dibiarkan kondisi tersebut nantinya akan berpengaruh pada hasil klasifikasi. Oleh karena itu selanjutnya akan dilakukan penanganan pada data yang tidak seimbang.

Tabel III
PERSENTASE UNTUK SETIAP KATEGORI PEUBAH PENJELAS

Peubah Penjelas	Kategori	Persentase Total (%)	Persentase Tidak Lulus (%)
Jenis Kelamin	Laki-laki	41.93	5.80
	Perempuan	58.07	5.17
Status Perkawinan	Belum Menikah	46.01	5.21
	Menikah	53.99	5.76
Status Penerimaan	Biasa	71.82	6.18
	Percobaan	24.02	4.52
	Fast Track	3.66	0.18
	PMDSU	0.50	0.08
Status PT Asal	PT Negeri	88.31	8.42
	PT Swasta	11.53	2.52
	PT Luar Negeri	0.16	0.02
Sumber Biaya	Beasiswa	56.29	3.49
Pendidikan	Sendiri	43.71	7.47
Jenis Pekerjaan	Instansi Negeri	24.12	2.30
	Instansi Swasta	17.73	2.52
	Luar Negeri	0.06	0.02
	Tidak Bekerja	58.09	6.12
Program Studi	Sains	83.01	8.85
	Sosial	16.99	2.12

Persentase setiap kategori untuk peubah penjelas dengan tipe kategorik ditunjukkan pada Tabel 3. Peubah penjelas dengan tipe numerik yang digunakan pada penelitian ini adalah IPK asal S1 dan usia masuk SPs-IPB. Gambar 1(a) menunjukkan *boxplot* untuk peubah usia. Nilai median usia masuk SPs-IPB untuk mahasiswa yang tidak lulus lebih tinggi dibandingkan mahasiswa yang lulus. Mahasiswa yang lulus memiliki usia minimum 19 tahun dan mahasiswa yang tidak lulus memiliki usia minimum 20 tahun sedangkan kedua kategori memiliki usia maksimum yang sama yaitu 56 tahun. Hal ini menunjukkan bahwa mahasiswa yang masuk

SPs-IPB baik pada usia muda maupun usia tua dapat berpotensi tidak lulus.



Gambar 1. *Boxplot* untuk peubah penjelas (a) usia masuk SPs-IPB dan (b) IPK asal S1

Gambar 1(b) menunjukkan *boxplot* untuk peubah IPK asal S1. Terlihat bahwa mahasiswa yang lulus memiliki kotak *boxplot* yang lebih lebar yang menunjukkan bahwa nilai IPK lebih beragam. Kemudian nilai median untuk kategori lulus lebih besar dibandingkan kategori tidak lulus. Kedua kategori memiliki nilai IPK minimum yang sama yaitu 2.00 sedangkan IPK maksimum untuk mahasiswa yang lulus adalah 4.00 dan mahasiswa yang tidak lulus adalah 3.94. Hal ini menunjukkan bahwa walaupun seorang mahasiswa memiliki IPK asal yang tinggi tetap dapat berpotensi tidak lulus, begitu juga sebaliknya.

Selajutnya akan dibahas persentase mahasiswa tidak lulus setiap kategori peubah penjelas yang juga tersedia pada Tabel 3. Secara eksploratif karakteristik mahasiswa yang tidak lulus dapat dilihat dari besar persentase kategori pada setiap peubah penjelas. Dilihat dari status penerimaannya, mahasiswa dengan status biasa memiliki persentase tidak lulus yang paling besar. Lalu untuk status PT asal, status PT negeri memiliki persentase mahasiswa tidak lulus terbesar. Selanjutnya berdasarkan sumber biaya pendidikan, mahasiswa yang tidak lulus didominasi oleh mahasiswa yang membiayai pendidikannya sendiri. Kemudian jika dilihat dari segi pekerjaan, mahasiswa yang tidak bekerja dan berstatus tidak lulus memiliki persentase terbesar. Peubah selanjutnya yaitu program studi. Persentase mahasiswa tidak lulus pada program studi sains jauh lebih besar dibandingkan program studi sosial. Peubah jenis kelamin dan status perkawinan memiliki persentase mahasiswa tidak lulus yang tidak jauh berbeda antar kategorinya. Mahasiswa yang lulus dan tidak lulus memiliki usia masuk SPs-IPB dan IPK asal yang tidak jauh berbeda.

B. Kinerja Klasifikasi SVM

Sebelum dilakukan pemodelan SVM, data terlebih dahulu dibagi menjadi 80% data latih dan 20% data uji. Pembagian tersebut diatur sedemikian rupa sehingga perbandingan kelas minoritas dan kelas mayoritas pada data latih dan data uji relatif sama dengan data asli. Hal ini dilakukan agar dapat merepresentasikan kondisi ketidakseimbangan pada data awal. Jenis SVM yang digunakan untuk pemodelan yaitu SVM linier dan SVM non-linier menggunakan fungsi kernel polinomial dan RBF.

Hasil kinerja klasifikasi pada ketiga jenis SVM menghasilkan nilai akurasi, sensitivitas, dan spesifisitas yang sama untuk setiap ulangannya yaitu berturut-turut sebesar 89.08%, 0.00%, dan 100.00%. Kemampuan SVM mengklasifikasikan mahasiswa secara tepat ditunjukkan oleh nilai akurasi yang memiliki nilai yang cukup besar yaitu 89.08%. Selain itu SVM pun dapat mengklasifikasikan semua mahasiswa yang berstatus lulus dengan tepat yang ditunjukkan dengan nilai spesifisitas sebesar 100.00%. Namun ternyata nilai sensitivitasnya adalah 0.00%, yang menunjukkan bahwa tidak ada satupun mahasiswa yang berstatus tidak lulus yang diklasifikasikan dengan tepat. Hal ini akan merugikan SPs-IPB jika menerima mahasiswa yang sebenarnya tidak lulus. Oleh karena itu, selanjutnya akan dilakukan penanganan pada ketidakseimbangan data untuk meningkatkan kemampuan SVM dalam mengklasifikasikan mahasiswa tidak lulus.

C. Kinerja Klasifikasi SVM dengan SMOTE

Rendahnya nilai sensitivitas hasil klasifikasi pada data yang tidak seimbang mengindikasikan perlunya dilakukan penanganan. Metode SMOTE digunakan untuk mengatasi ketidakseimbangan pada data keberhasilan studi mahasiswa. Berdasarkan hasil SMOTE diperoleh banyaknya mahasiswa lulus dan tidak lulus yang relatif seimbang. Tabel 4 menunjukkan komposisi pembagian data latih setelah melalui tahap SMOTE. Hasil rataan kinerja klasifikasi pada data latih yang sudah melalui tahap SMOTE ditunjukkan pada Tabel 5. Terlihat bahwa rataan nilai sensitivitasnya mengalami

peningkatan baik pada SVM linier, polinomial maupun RBF dengan nilai berturut-turut 45.81%, 47.98%, dan 44.84%. Hal ini menunjukkan bahwa penanganan ketidakseimbangan data menggunakan SMOTE berhasil meningkatkan kemampuan SVM dalam mengklasifikasikan mahasiswa tidak lulus. Selain itu terlihat bahwa nilai rata-rata akurasi dan spesifisitas tertinggi dihasilkan dari jenis SVM RBF. Namun nilai rata-rata sensitivitas tertinggi dihasilkan dari SVM polinomial. Oleh karena itu dapat dikatakan bahwa jenis SVM yang tepat digunakan untuk mengklasifikasikan keberhasilan studi mahasiswa adalah SVM RBF.

Tabel IV
KOMPOSISI PEMBAGIAN DATA LATIH SEBELUM DAN SESUDAH MELALUI TAHAP SMOTE

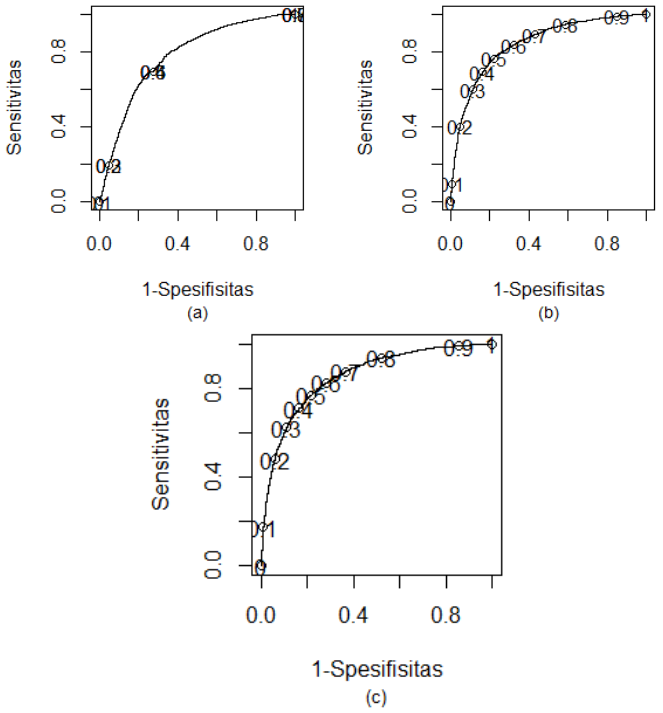
Kelas	Data Latih	Data Latih Hasil SMOTE
Minoritas	435 (10.98%)	3480 (46.97%)
Mayoritas	3527 (89.02%)	3527 (50.33%)
Total	3962 (100.00%)	7007 (100.00%)

Tabel V
RATAAN KINERJA KLASIFIKASI BERBAGAI JENIS SVM SETELAH MELALUI TAHAP SMOTE

Jenis SVM	Kinerja Klasifikasi (%)		
	Akurasi	Sensitivitas	Spesifisitas
Linier	70.42	45.81	73.43
Polinomial	70.83	47.98	73.63
RBF	73.97	44.84	77.54

D. Kinerja Klasifikasi SVM dengan SMOTE menggunakan Cut Off Terbaik

Setelah dilakukan penanganan data tidak seimbang menggunakan SMOTE, nilai sensitivitas yang dihasilkan mengalami peningkatan namun dapat dikatakan masih rendah. Nilai tersebut dapat ditingkatkan lagi dengan mengubah nilai *cut off* yang digunakan. Gambar 2 menunjukkan kurva ROC untuk SVM linier, polinomial, dan RBF beserta nilai *cut off* untuk masing-masing jenis SVM. Kinerja klasifikasi yang baik dihasilkan ketika nilai *cut off* sensitivitas tinggi dan nilai 1-spesifisitas rendah. Nilai *cut off* tersebut untuk ketiga jenis SVM ditunjukkan pada Tabel 6. Selanjutnya dilakukan klasifikasi kembali pada data latih hasil SMOTE menggunakan nilai *cut off* tersebut untuk kemudian dibandingkan kinerja klasifikasinya untuk ditentukan nilai *cut off* yang memberikan kinerja terbaik.



Gambar 2. Kurva ROC untuk jenis SVM (a) linier, (b) polinomial, dan (c) RBF beserta nilai *cut off* untuk masing-masing jenis SVM

Tabel 6 menunjukkan hasil rataan kinerja klasifikasi SVM pada berbagai nilai *cut off*. Terlihat bahwa semakin besar nilai *cut off* semakin besar pula nilai sensitivitasnya. Namun hal tersebut juga diiringi dengan penurunan nilai akurasi dan spesifisitasnya. Penentuan nilai *cut off* yang akan digunakan didasarkan pada ketiga ukuran keakuratan tersebut. Nilai sensitivitas diharapkan dapat meningkat dibandingkan ketika menggunakan nilai *cut off default* yaitu 0.5 serta memiliki akurasi dan spesifisitas yang tidak terlalu rendah. SVM linier, polinomial, dan RBF memiliki rataan kinerja klasifikasi paling baik pada *cut off* 0.6.

Tabel 7 menunjukkan hasil rataan kinerja klasifikasi SVM setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik. Terlihat bahwa nilai rataan akurasi dan spesifisitas tertinggi dihasilkan dari jenis SVM linier yaitu berturut-turut sebesar 68.80% dan 71.44%. Namun nilai rataan sensitivitas tertinggi dihasilkan dari SVM polinomial

yaitu sebesar 55.24%. Sehingga dapat dikatakan bahwa SVM linier memberi hasil yang lebih baik. Tetapi berdasarkan Gambar 4, terlihat bahwa terdapat banyak pencilan pada *boxplot* nilai akurasi dan spesifisitas SVM linier sedangkan boxplot sensitivitasnya memiliki kotak yang lebar yang menunjukkan bahwa SVM linier memiliki ragam yang cukup besar. Oleh karena itu, jenis SVM linier tidak dapat dipilih. SVM RBF memiliki rataan nilai akurasi dan spesifisitas yang sedikit lebih rendah dari SVM linier serta rataan nilai sensitivitas yang sedikit lebih rendah dari SVM polinomial sehingga dapat dipilih sebagai jenis SVM yang yang tepat untuk mengklasifikasikan keberhasilan studi mahasiswa.

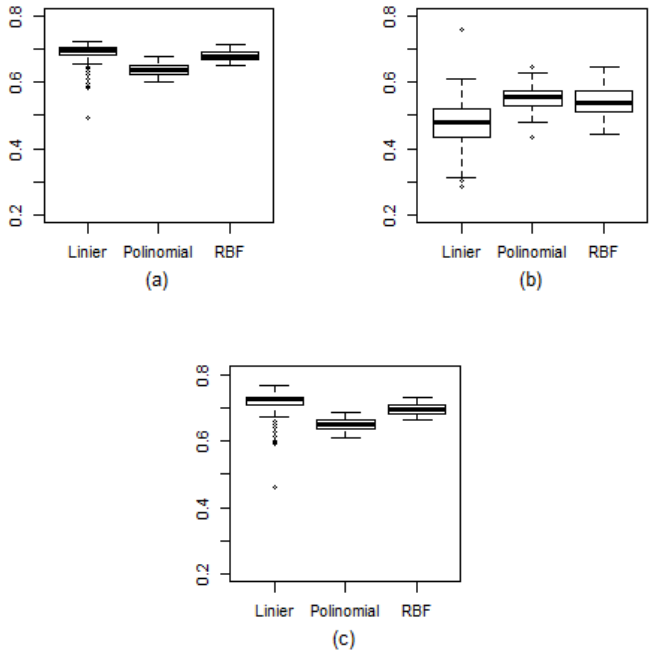
Tabel VI
RATAAN KINERJA KLASIFIKASI SVM PADA BERBAGAI JENIS SVM DAN BERBAGAI NILAI CUT OFF

Cut Off	Kinerja Klasifikasi (%)		
	Akurasi	Sensitivitas	Spesifisitas
Linier			
0.4	70.97	44.00	74.28
0.5	70.00	45.09	73.06
0.6	68.80	47.26	71.44
Polinomial			
0.4	77.81	37.33	82.77
0.5	71.95	46.90	75.03
0.6	63.82	55.24	64.88
0.7	55.27	63.44	54.27
RBF			
0.3	82.26	27.63	88.96
0.4	78.06	35.93	83.23
0.5	73.16	46.06	76.48
0.6	67.92	54.14	69.60
0.7	61.16	63.53	60.88

Tabel VII
RATAAN KINERJA KLASIFIKASI BERBAGAI JENIS SVM SETELAH MELALUI TAHAP SMOTE DAN MENGGUNAKAN CUT OFF TERBAIK

Jenis SVM	Kinerja Klasifikasi (%)		
	Akurasi	Sensitivitas	Spesifisitas
Linier	68.80	47.26	71.44
Polinomial	63.82	55.24	64.88
RBF	67.92	54.14	69.60

Tabel 8 menunjukkan rataan nilai sensitivitas sebelum dan sesudah dilakukan penanganan. Terlihat bahwa pemodelan SVM pada data asli menghasilkan rataan nilai sensitivitas yang sangat rendah yaitu 0.00%, yang menunjukkan bahwa tidak ada satupun mahasiswa yang berstatus tidak lulus yang dapat diklasifikasikan dengan tepat. Kemudian setelah melalui tahap SMOTE, sensitivitas ketiga jenis SVM berhasil ditingkatkan menjadi lebih dari 40%. Selanjutnya setelah melalui tahap SMOTE dan menggunakan nilai *cut off* terbaik, sensitivitas berhasil ditingkatkan kembali.



Gambar 3. *Boxplot* nilai (a) akurasi, (b) sensitivitas, dan (c) spesifisitas berbagai jenis SVM setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik

Tabel VIII
PERBANDINGAN RATAAN NILAI SENSITIVITAS BERBAGAI JENIS SVM

Jenis SVM	Sensitivitas (%)		
	Tanpa SMOTE	SMOTE	SMOTE + *cut off* terbaik
Linier	0.00	45.81	47.26
Polinomial	0.00	47.98	55.24
RBF	0.00	44.84	54.14

V. SIMPULAN

Pemodelan SVM pada data tidak seimbang menghasilkan kinerja yang kurang baik dengan nilai

sensitivitas 0.00%. Penanganan data tidak seimbang menggunakan SMOTE berhasil meningkatkan kinerja klasifikasi SVM dalam mengklasifikasikan mahasiswa yang tidak lulus. Jenis SVM yang tepat digunakan untuk melakukan pemodelan keberhasilan studi mahasiswa adalah SVM RBF. Ketika menggunakan nilai *cut off* terbaik dari masing-masing jenis SVM, nilai sensitivitas berhasil untuk ditingkatkan kembali. SVM RBF masih memberi hasil yang paling baik ketika menggunakan *cut off* 0.6.

DAFTAR PUSTAKA

Agwil, W. (2015). Pengklasifikasian status infeksi hookworm pada kucing dengan menggunakan smote support vector machine dan boosting support vector machine [tesis]. *Bogor(ID):Institut Pertanian Bogor*.

Bekkar, M., H. K. Djemaa, and T. A. Alitouche (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications 3*, 27–38.

Bunkhumpornpat, C., K. Sinapiromsaran, and C. Lursinsap (2012). Dbsmote: density-based synthetic minority over-sampling technique. *Application Intelligence 36*, 664–684.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Han, H., W. Y. Wang, and B. H. Mao (2005). Borderline-smote: a new over-sampling method in imbalance data sets learning. *Springer-Verlag Berlin Heidelberg*, 878–887.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Permatasari, I. (2009). Kajian performa program studi magister sekolah pascasarjana ipb [skripsi]. *Bogor(ID):Institut Pertanian Bogor*.

Scolkopf, B. and A. J. Smola (2000). *Learning with Kernels*. Massachusetts: MIT Press.

Wang, B. X. and N. Japkowicz (2010). Boosting support vector machines for imbalanced data sets. *Knowledge Information System 25*, 1–20.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. e. a. Yu (2008). Top 10 algorithms in data mining. *Knowledge Information System 14*, 1–37.