PEMODELAN *SUPPORT VECTOR MACHINE* DATA BERKATEGORI TIDAK SEIMBANG BESERTA PENANGANANNYA DENGAN *SMOTE*
(Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015)

OCTAVIA DWI AMELIA

DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2018

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi berjudul Pemodelan *Support Vector Machine* Data Berkategori Tidak Seimbang beserta Penanganannya dengan *SMOTE* (Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015) adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2018

Octavia Dwi Amelia
NIM G14140078

ABSTRAK

OCTAVIA DWI AMELIA. Pemodelan *Support Vector Machine* Data Berkategori Tidak Seimbang beserta Penanganannya dengan *SMOTE* (Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015). Dibimbing oleh AGUS MOHAMAD SOLEH dan SEPTIAN RAHARDIANTORO.

Dalam upaya mempertahankan reputasi Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB), perlu diterapkan sistem penerimaan mahasiswa baru pada program magister dengan lebih selektif. Penelitian ini bertujuan untuk memodelkan kategori status kelulusan mahasiswa SPs-IPB berdasarkan karakteristik dan latar belakang pendidikan mahasiswa. Pada data tersebut, ditelusuri bahwa banyaknya mahasiswa dengan kategori tidak lulus (10.97%) jauh lebih sedikit dibandingkan banyaknya mahasiswa dengan kategori lulus (89.03%). Hal ini menunjukkan bahwa terdapat ketidakseimbangan kategori pada data. Pemodelan *Support Vector Machine* (SVM) menggunakan SVM linier, polinomial, dan *radial basis function* (RBF) belum mampu memprediksi kategori minoritas, yaitu mahasiswa yang berpotensi tidak lulus ditunjukkan dengan nilai sensitivitas 0.00%. Oleh karena itu, penelitian ini menerapkan metode *Sythetic Minority Oversampling Technique* (SMOTE) dalam rangka menangani ketidakseimbangan kategori pada data, yang selanjutnya diterapkan juga ketika menggunakan nilai *cut off* terbaik dari setiap jenis SVM. Berdasarkan evalusi yang dilakukan, model SVM RBF dengan *cut off* 0.6 mampu memprediksi mahasiswa pada kategori minoritas ditunjukkan dengan nilai sensitivitas yang mencapai 54.14%.

Kata kunci: kategori tidak seimbang, sensitivitas, SMOTE, SPs-IPB, SVM

ABSTRACT

OCTAVIA DWI AMELIA. Support Vector Machine Modeling for Imbalanced Data and Its Handling with SMOTE (Case Study: The Success of IPB Master Program Student in 2011-2015). Supervised by AGUS MOHAMAD SOLEH and SEPTIAN RAHARDIANTORO.

In order to maintain the reputation of Graduate School of Bogor Agricultural University (SPs-IPB), a more selective admissions system in the master program needs to be applied. This study aims to model the graduate status of SPs-IPB students based on characteristics and educational background of the students. In the data, it is observed that the number of students with non-graduate category (10.97%) is much lower than the graduate category (89.03%). This indicates that there is an imbalance category in the data. Support Vector Machine (SVM) modeling using SVM linear, polynomial, and radial base function (RBF) has not been able to predict the minority category, ie students who potentially fail to graduate indicated by sensitivity value of 0.00%. Therefore, this study applies Sythetic Minority Oversampling Technique (SMOTE) in order to handle imbalanced data problem, which is further also applied when using the best cut off value of each type of SVM. Based on the evaluation, the result shows that SVM RBF model with cut off 0.6 can predict students in minority category which is indicated by sensitivity value reaching 54.14%.

Keywords: imbalanced data, sensitivity, SMOTE, SPs-IPB, SVM

PEMODELAN *SUPPORT VECTOR MACHINE* DATA BERKATEGORI TIDAK SEIMBANG BESERTA PENANGANANNYA DENGAN *SMOTE* (Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015)

OCTAVIA DWI AMELIA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Statistika
pada
Departemen Statistika

DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2018

Judul Skripsi: Pemodelan *Support Vector Machine* Data Berkategori Tidak Seimbang beserta Penanganannya dengan *SMOTE* (Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015)
Nama : Octavia Dwi Amelia
NIM : G14140078

Disetujui oleh

Dr Agus Mohamad Soleh, MT
Pembimbing I

Septian Rahardiantoro, MSi
Pembimbing II

Diketahui oleh

Dr Anang Kurnia, MSi
Ketua Departemen

Tanggal Lulus: 31 JUL 2018

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian ini adalah pemodelan klasifikasi, dengan judul Pemodelan *Support Vector Machine* Data Berkategori Tidak Seimbang beserta Penanganannya dengan *SMOTE* (Studi Kasus: Keberhasilan Studi Mahasiswa Program Magister IPB Tahun 2011 sampai 2015).

Terima kasih penulis ucapkan kepada Bapak Dr Agus Mohamad Soleh, MT dan Bapak Septian Rahardiantoro, MSi selaku dosen pembimbing yang telah membimbing dan banyak memberi saran dalam proses penyusunan karya ilmiah ini. Kemudian terima kasih kepada Ibu Dr Utami Dyah Syafitri, MSi selaku dosen penguji yang juga telah memberikan saran kepada penulis. Selain itu, terima kasih juga penulis ucapkan pada Bapak Pungki Prayughi, SKom, MKom selaku Kepala Bagian Informasi TU Sekolah Pascasarjana IPB yang telah membantu dalam memperoleh data. Tak lupa penulis juga mengucapkan terima kasih kepada seluruh dosen dan pegawai tata usaha Departemen Statistika, teman sebimbingan, teman-teman statistika angkatan 51, keluarga Pondok Nuansa Sakinah dan juga kepada kedua orang tua serta keluarga atas segala dukungan, bantuan, doa dan kasih sayangnya.

Semoga karya ilmiah ini bermanfaat.

Bogor, Juli 2018

Octavia Dwi Amelia

DAFTAR ISI

DAFTAR TABEL x

DAFTAR GAMBAR x

PENDAHULUAN 1

Latar Belakang 1

Tujuan Penelitian 2

TINJAUAN PUSTAKA 2

Ketidakseimbangan Kategori pada Data 2

Synthetic Minority Oversampling Technique 3

Support Vector Machine 4

Kinerja Klasifikasi 6

METODE 7

Data 7

Metode Penelitian 8

HASIL DAN PEMBAHASAN 9

Karakteristik Mahasiswa Program Magister IPB 9

Kinerja Klasifikasi SVM 13

Kinerja Klasifikasi SVM dengan SMOTE 15

Kinerja Klasifikasi SVM dengan SMOTE menggunakan *Cut Off* Terbaik 16

SIMPULAN 20

DAFTAR PUSTAKA 21

RIWAYAT HIDUP 22

DAFTAR TABEL

1 Matriks konfusi untuk peubah respon dengan 2 kategori 7
2 Daftar peubah penjelas yang digunakan 8
3 Komposisi pembagian data latih dan data uji pada masing-masing kategori data 13
4 Nilai parameter yang digunakan dalam melakukan pemodelan berbagai jenis SVM 13
5 Rataan kinerja klasifikasi pada data uji berbagai jenis SVM tanpa penanganan ketidakseimbangan kategori pada data 14
6 Komposisi pembagian data latih sebelum dan sesudah melalui tahap SMOTE 15
7 Rataan kinerja klasifikasi pada data uji berbagai jenis SVM setelah melalui tahap SMOTE 15
8 Rataan kinerja klasifikasi berbagai jenis SVM pada data uji dengan berbagai nilai *cut off* 18
9 Rataan kinerja klasifikasi pada data uji berbagai jenis SVM setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik 18
10 Perbandingan rataan nilai sensitivitas pada data uji berbagai jenis SVM pada 3 tahapan analisis 20
11 Kinerja klasifikasi SVM RBF dengan *cut off* 0.6 pada data keseluruhan setelah melalui tahap SMOTE 20

DAFTAR GAMBAR

1 Ilustrasi plot tebaran data yang dapat dipisahkan secara linier sempurna 4
2 Ilustrasi plot tebaran data yang tidak dapat dipisahkan secara linier sempurna 5
3 Ilustrasi plot tebaran data yang tidak dapat dipisahkan secara linier 6
4 *Pie chart* banyaknya mahasiswa berdasarkan status kelulusan 9
5 *Stack column chart* banyaknya mahasiswa pada setiap kategori peubah penjelas 10
6 *Boxplot* untuk peubah penjelas (a) usia masuk SPs-IPB dan (b) IPK asal S1 berdasarkan status kelulusan 11
7 *Stack bar chart* banyaknya mahasiswa pada setiap kategori peubah penjelas berdasarkan status kelulusan 12
8 Rataan nilai sensitivitas berbagai jenis SVM pada berbagai persentase kategori minoritas 14
9 *Boxplot* nilai (a) akurasi, (b) sensitivitas, dan (c) spesifisitas berbagai jenis SVM setelah melalui tahap SMOTE 16
10 Kurva ROC untuk jenis SVM (a) linier, (b) polinomial, dan (c) RBF beserta nilai *cut off* untuk masing-masing jenis SVM 17
11 *Boxplot* nilai (a) akurasi, (b) sensitivitas, dan (c) spesifisitas berbagai jenis SVM setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik 19

PENDAHULUAN

Latar Belakang

Sekolah Pascasarjana Institut Pertanian Bogor (SPs-IPB) didirikan pada tahun 1975 dan awalnya hanya terdiri dari 7 program studi (IPB 2013). Namun kini sudah berkembang menjadi 71 program studi magister dan 43 program studi doktor. Program studi yang ada di SPs-IPB sebagian besar sudah berakreditasi A. SPs-IPB memiliki standar akademik serta daya saing lulusan yang tinggi. Oleh karena itu, diperlukan penyusunan strategi dalam rangka mempertahankan reputasi SPs-IPB. Salah satu cara yang dapat dilakukan adalah dengan menerapkan sistem penerimaan mahasiswa baru yang lebih selektif. Hal ini dimaksudkan agar mahasiswa yang diterima adalah mahasiswa yang berpotensi besar untuk lulus dari SPs-IPB sehingga dapat mengurangi jumlah mahasiswa yang berpotensi tidak lulus. Oleh karena itu, kasus keberhasilan studi mahasiswa SPs-IPB khususnya program magister menjadi hal yang sangat menarik untuk dikaji.

Kajian mengenai keberhasilan studi mahasiswa sebelumnya sudah pernah dilakukan. Permatasari (2009) menggunakan regresi logistik biner dan regresi cox dalam mengkaji karakteristik mahasiswa yang berpengaruh pada tiga indikator keberhasilan studi mahasiswa program magister IPB yaitu lulus atau keluar, IPK akhir dan masa studi. Penelitian ini berupaya untuk mengkaji keberhasilan studi mahasiswa berdasarkan status kelulusan, yaitu lulus atau tidaknya mahasiswa dari SPs-IPB. Karena peubah respon yang digunakan adalah status kelulusan yang merupakan peubah kategorik, penelitian ini menerapkan pemodelan *Support Vector Machine* (SVM) yang merupakan salah satu metode klasifikasi yang paling berpengaruh dan banyak digunakan dalam *data mining*. Metode SVM diterapkan dengan mempertimbangkan karakteristik dan latar belakang pendidikan mahasiswa magister SPs-IPB. Metode ini dipilih karena dianggap mampu mengklasifikasikan data yang dipisahkan secara linier maupun non-linier. SVM menggunakan sebuah bidang pemisah dalam melakukan klasifikasi. Penentuan bidang pemisah tersebut didasarkan pada konsep *maximal margin classifier*. Hal ini yang menyebabkan SVM mampu menghasilkan kinerja klasifikasi yang sangat baik tidak hanya pada data latih namun juga pada data uji (Wu *et al*. 2008).

Pada praktiknya penelitian ini menggunakan metode SVM untuk diterapkan pada data keberhasilan studi mahasiswa SPs-IPB dengan memperhatikan adanya ketidakseimbangan kategori pada data. Kondisi data dikatakan berkategori tidak seimbang karena banyaknya mahasiswa yang tidak lulus jauh lebih sedikit dibandingkan banyaknya mahasiswa yang menempuh pendidikan hingga lulus. Umumnya kinerja klasifikasi dari berbagai metode klasifikasi akan mengalami penurunan bila diterapkan pada data yang memiliki kategori tidak seimbang. Salah satu metode klasifikasi yang kinerjanya menurun ketika diterapkan pada kondisi tersebut yaitu metode SVM (Ganganwar 2012). Ketidakseimbangan kategori pada data dapat mengakibatkan salah klasifikasi pada kategori minoritas, yaitu kategori dengan banyak amatan yang jauh lebih sedikit. Hal ini tentunya akan merugikan SPs-IPB jika menerima mahasiswa yang berpotensi besar tidak dapat menyelesaikan proses studinya. Oleh karena itu, penelitian ini menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) dalam upaya menangani

kondisi kategori pada data yang tidak seimbang. Konsep dasar SMOTE ialah dengan menambahkan data buatan pada kategori minoritas sehingga membentuk data baru yang memiliki kategori lebih seimbang. Penggunaan metode SMOTE ini dipilih, karena didukung oleh penelitian yang dilakukan oleh Agwil (2015) pada pengklasifikasian status infeksi *hookworm* pada kucing yang menunjukkan bahwa SMOTE dapat meningkatkan kinerja klasifikasi SVM dalam mengklasifikasikan kategori minoritasnya. Singkatnya, penelitian ini berupaya untuk mengkaji keberhasilan studi mahasiswa magister SPs-IPB dengan menggunakan metode SVM serta memperhatikan kondisi ketidakseimbangan kategori pada data yang diatasi dengan metode SMOTE.

Tujuan Penelitian

Tujuan dari penelitian ini adalah menerapkan dan mengevaluasi prediksi keberhasilan studi mahasiswa program magister SPs-IPB dengan menggunakan metode SVM dengan dan tanpa menggunakan metode SMOTE.

TINJAUAN PUSTAKA

Pada bagian ini, akan disajikan beberapa teori dasar mengenai konsep-konsep dan metode analisis yang digunakan dalam penelitian ini. Dimulai dari konsep ketidakseimbangan kategori pada data, metode SMOTE, SVM dan beberapa jenisnya, serta konsep mengenai kinerja klasifikasi.

Ketidakseimbangan Kategori pada Data

Ketidakseimbangan kategori pada data terjadi ketika suatu kategori memiliki banyak amatan yang jauh lebih sedikit dibandingkan kategori lainnya (Ganganwar 2012). Ketika menggunakan dua kategori, kategori dengan amatan yang sedikit disebut kategori minoritas, sedangkan kategori lainnya disebut kategori mayoritas. Ketidakseimbangan kategori pada data perlu diperiksa terlebih dahulu sebelum melakukan analisis klasifikasi. Analisis klasifikasi secara umum kurang memadai dalam mengatasi ketidakseimbangan kategori pada data karena algoritme dibuat tanpa memperhatikan hal tersebut (Han *et al*. 2005). Akibatnya kategori minoritas akan mengalami salah klasifikasi. Oleh karena itu, perlu dilakukan penanganan pada kondisi ketidakseimbangan tersebut. Usaha yang dapat dilakukan dalam menangani kasus ini terbagi menjadi dua, yaitu penanganan pada tingkat algoritme dan tingkat data (Han *et al*. 2005). Penanganan pada tingkat algoritme dilakukan dengan memodifikasi algoritme yang sudah ada ataupun membuat algoritme baru. Penanganan pada tingkat data dilakukan dengan menggunakan metode *resampling*. Menurut Ganganwar (2012), metode resampling yang dapat digunakan yaitu *undersampling* dan *oversampling*. *Undersampling* dilakukan dengan mengurangi amatan pada kategori mayoritas sedangkan *oversampling* dilakukan dengan

menambah amatan pada kategori minoritas. Namun, kedua metode *resampling* ini memiliki kekurangan. *Undersampling* dapat mengakibatkan terjadinya kehilangan informasi sedangkan *oversampling* dapat menimbulkan masalah *overfitting*.

Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode *oversampling* dengan menciptakan data buatan (sintetik) pada kategori minoritas untuk mengatasi masalah ketidakseimbangan kategori pada data (Chawla *et al*. 2002). Metode SMOTE dapat mengatasi masalah *overfitting* yang ditimbulkan ketika melakukan *oversampling* dengan pengembalian (Kotsiantis *et al.* 2006). *Oversampling* dengan metode SMOTE dilakukan dengan memanfaatkan konsep *k*-tetangga terdekat. Tahapan dalam penerapan metode SMOTE adalah sebagai berikut.

1. Menghitung jarak antar amatan pada kategori minoritas.
   Jarak dihitung menggunakan jarak Euclid. Ketika data terdiri dari tipe numerik dan kategorik, perhitungan jarak tetap menggunakan jarak Euclid namun dengan menggunakan nilai median simpangan baku peubah numerik sebagai selisih nilai peubah kategorik. Nilai median ini dihitung ketika kategori antar amatan berbeda. Menurut Raykov dan Marcoulides (2008) jarak Euclid ditunjukkan oleh persamaan berikut
   $$D(\boldsymbol{a},\boldsymbol{b}) = \sqrt{(\boldsymbol{a}-\boldsymbol{b})'(\boldsymbol{a}-\boldsymbol{b})} = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2} \qquad (1)$$
   dengan $D(\boldsymbol{a},\boldsymbol{b})$ adalah jarak antara vektor amatan $\boldsymbol{a} = (a_1, a_2, \dots, a_p)'$ dan vektor amatan $\boldsymbol{b} = (b_1, b_2, \dots, b_p)'$, a_l adalah nilai amatan a pada peubah ke-l, y_l adalah nilai amatan b pada peubah ke-l, dan p adalah banyaknya peubah.
2. Menentukan nilai *k* untuk *k*-tetangga terdekat.
3. Menentukan k-tetangga terdekat untuk amatan terpilih dengan mengurut jarak antara amatan terpilih dengan semua amatan kategori minoritas.
4. Memilih secara acak 1 amatan dari k-tetangga terdekatnya.
5. Melakukan perhitungan untuk membangkitkan data buatan dengan prosedur sebagai berikut.
   a. Data Numerik
      i. Mengalikan jarak antara amatan terpilih dan tetangga terdekat yang terpilih pada tahap 4 dengan bilangan acak antara 0 sampai 1.
      ii. Menambahkan hasil perkalian tersebut dengan amatan terpilih sehingga diperoleh amatan baru.
   b. Data Kategorik
      i. Menentukan kategori yang paling sering muncul pada amatan terpilih dan k-tetangga terdekatnya. Jika nilainya sama maka dipilih secara acak.
      ii. Jadikan kategori tersebut sebagai amatan baru.
6. Mengulangi tahap 3 sampai 5 hingga banyaknya amatan kategori minoritas dan mayoritas relatif seimbang.

Support Vector Machine (SVM)

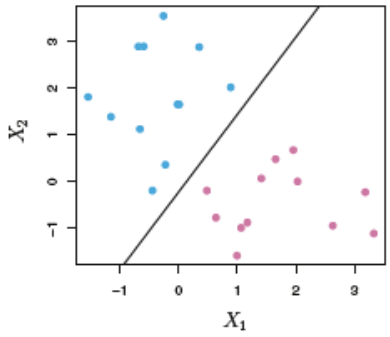
Support vector machine (SVM) merupakan metode klasifikasi yang menggunakan bidang pemisah (*hyperplane*) dalam melakukan klasifikasi pada data latih dan selanjutnya diharapkan dapat mengklasifikasikan data uji dengan tepat. Pada ruang berdimensi p, bidang pemisah akan berdimensi $p-1$. Klasifikasi dapat dilakukan menggunakan berbagai kemungkinan bidang pemisah, sehingga perlu ditentukan bidang pemisah yang terbaik. Bidang pemisah terbaik ditentukan menggunakan *maximal margin classifier*, yaitu dengan menentukan bidang pemisah yang memiliki *margin* paling besar (*maximal margin hyperplane*). *Margin* adalah jarak terdekat amatan data latih terhadap bidang pemisah. Amatan yang memiliki jarak terdekat dengan bidang pemisah disebut *support vector* (James *et al.* 2013).

SVM linier

SVM linier diterapkan pada data yang dapat dipisahkan secara linier. Data yang dapat dipisahkan secara linier terbagi menjadi dua yaitu dapat dipisahkan secara sempurna dan tidak dapat dipisahkan secara sempurna. Gambar 1 menunjukkan contoh data yang dapat dipisahkan menggunakan bidang pemisah secara sempurna. Menurut Schölkopf dan Smola (2002), secara umum fungsi keputusan untuk jenis SVM linier dapat dituliskan seperti persamaan berikut

$$f(\boldsymbol{x}) = \begin{cases} -1, & \boldsymbol{w}'\boldsymbol{x} + b < 0 \\ +1, & \boldsymbol{w}'\boldsymbol{x} + b > 0, \end{cases} \tag{2}$$

dengan $f(\boldsymbol{x})$ merupakan prediksi kategori untuk vektor $\boldsymbol{x}$ dengan nilai -1 menunjukkan kategori negatif dan $+1$ menunjukkan kategori positif, $\boldsymbol{w}$ adalah vektor berukuran $p \times 1$ yang tegak lurus dengan bidang pemisah dengan p menunjukkan banyak peubah penjelas, $\boldsymbol{x}$ adalah vektor berukuran $p \times 1$ yang merupakan vektor amatan yang ingin diprediksi kategorinya dan b adalah konstanta.



Gambar 1 Ilustrasi plot tebaran data yang dapat dipisahkan secara linier sempurna

Bidang pemisah terbaik ditentukan dengan meminimumkan

$$\frac{1}{2}\|\boldsymbol{w}\|^2 \tag{3}$$

dengan konstrain

$$y_i(\boldsymbol{w}'\boldsymbol{x_i} + b) \geq 1. \tag{4}$$

y_i adalah kategori peubah respon pada data latih $\boldsymbol{x_i}$ dengan $y_i \in \{-1, +1\}$, $i = 1,2,\ldots,m$. Persoalan di atas dapat diselesaikan dengan menggunakan pengali *Langrange* berikut

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i (y_i(\boldsymbol{w}'\boldsymbol{x_i} + b) - 1) \ , \ \ \alpha_i \geq 0 \qquad (5)$$

dengan α_i adalah konstanta *lagrange*. Persamaan (5) selanjutnya diturunkan terhadap b dan $\boldsymbol{w}$ sehingga diperoleh hasil sebagai berikut

$$\frac{\partial}{\partial b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad (6)$$

$$\frac{\partial}{\partial \boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}. \qquad (7)$$

Persamaan (6) dan (7) dapat disubstitusikan ke dalam persamaan (5) menghasilkan persamaan

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}' \boldsymbol{x_j} \qquad (8)$$

yang selanjutnya dimaksimumkan dengan konstrain

$$\alpha_i \geq 0, i = 1, \ldots, m \text{ dan } \sum_{i=1}^{m} \alpha_i y_i = 0. \qquad (9)$$

Kemudian nilai α_i untuk setiap data latih akan diperoleh. Data latih yang terletak pada *margin* dengan nilai $\alpha_i > 0$ merupakan *support vector*. Sedangkan data latih dengan nilai $\alpha_i = 0$ tidak digunakan dalam fungsi keputusan. Dengan demikian fungsi keputusan yang hanya dipengaruhi *support vector* dapat dituliskan sebagai berikut

$$f(x) = \begin{cases} -1, \ \sum_{r=1}^{ns} \alpha_r y_r \boldsymbol{x}' \boldsymbol{x}_r + b < 0 \\ +1, \ \sum_{r=1}^{ns} \alpha_r y_r \boldsymbol{x}' \boldsymbol{x}_r + b > 0 \end{cases} \qquad (10)$$

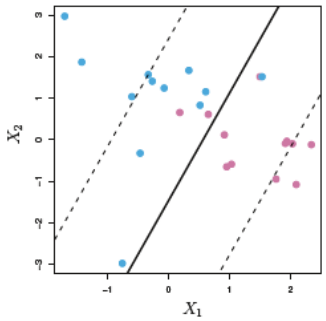
keterangan :

$\boldsymbol{x}$ = data uji

$\boldsymbol{x_r}$ = data yang merupakan *support vector*, $r = 1,2, \ldots, ns$

ns = banyak data yang merupakan *support vector*.

Gambar 2 menunjukkan contoh data yang tidak dapat dipisahkan menggunakan bidang pemisah secara sempurna. *Soft margin classifier* dapat digunakan untuk mengatasi hal tersebut. Metode ini menggunakan *soft margin hyperplane* yang membolehkan pelanggaran saat klasifikasi. Amatan mungkin berada pada sisi *margin* yang salah atau bahkan pada sisi bidang pemisah yang salah. Pada kasus seperti ini, keadaan tersebut tidak bisa dihindari (James *et al.* 2013).



Gambar 2 Ilustrasi plot tebaran data yang tidak dapat dipisahkan secara linier sempurna

Menurut Schölkopf dan Smola (2002), pada s*oft margin classifier* peubah *slack* (ξ) ditambahkan ke dalam fungsi sehingga bidang pemisah terbaik diperoleh dengan meminimumkan

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \xi_i \qquad (11)$$

dengan konstrain

$$y_i(\boldsymbol{w}'\boldsymbol{x_i} + b) \geq 1 - \xi_i, \xi_i \geq 0. \quad (12)$$

C merupakan parameter bernilai non-negatif yang menunjukkan penalti akibat pelanggaran saat klasifikasi, baik pelanggaran terhadap *margin* maupun bidang pemisah (James *et al.* 2013).

SVM non-linier

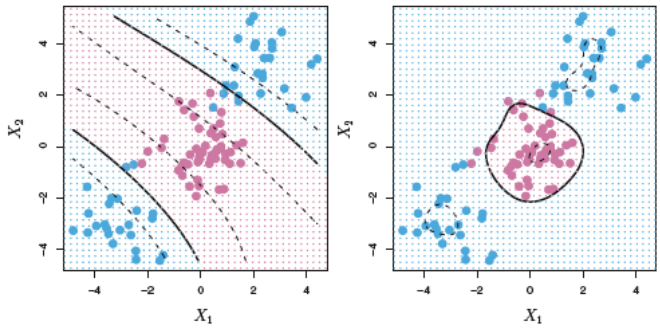
Selain dua kondisi data di atas, terdapat data yang tidak dapat dipisahkan secara linier sempurna maupun menggunakan *soft margin classifier*, seperti yang terlihat pada Gambar 3. Klasifikasi pada data seperti ini dapat diatasi dengan memperbesar ruang fitur menggunakan fungsi Kernel $K(\boldsymbol{x}, \boldsymbol{x_r})$. Menurut Schölkopf dan Smola (2002) fungsi keputusan pada persamaan (10) berubah menjadi persamaan berikut

$$f(x) = \begin{cases} -1, \ \sum_{r=1}^{ns} \alpha_r y_r K(\boldsymbol{x}, \boldsymbol{x_r}) + b < 0 \\ +1, \ \sum_{r=1}^{ns} \alpha_r y_r K(\boldsymbol{x}, \boldsymbol{x_r}) + b > 0. \end{cases} \quad (13)$$

Berikut ini merupakan beberapa fungsi Kernel :

1. Polinomial dengan derajat d : $K(\boldsymbol{x}, \boldsymbol{x_r}) = (1 + \boldsymbol{x}'\boldsymbol{x_r})^d$
2. *Radial Basis Function* (RBF) : $K(\boldsymbol{x}, \boldsymbol{x_r}) = exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x_r}\|^2)$

dengan d adalah derajat polinomial berupa bilangan bulat positif dan γ adalah suatu konstanta positif.



Gambar 3 Ilustrasi plot tebaran data yang tidak dapat dipisahkan secara linier

Kinerja Klasifikasi

Evaluasi kinerja klasifikasi dilakukan menggunakan matriks konfusi. Jika klasifikasi yang dilakukan pada peubah respon yang memiliki dua kategori, maka matriks konfusi memiliki ukuran 2×2 seperti yang ditunjukkan pada Tabel 1. Ketika terdapat ketidakseimbangan kategori pada data, kategori minoritas diberi label positif dan kategori mayoritas diberi label negatif (Bekkar *et al*. 2013). Sehingga pada penelitian ini kategori tidak lulus adalah kategori positif sedangkan kategori lulus adalah kategori negatif. TP menunjukkan banyaknya amatan kategori positif yang diprediksi kategori positif. FP menunjukkan banyaknya amatan kategori negatif yang diprediksi kategori positif. FN menunjukkan banyaknya amatan kategori positif yang diprediksi kategori negatif. TN menunjukkan banyaknya amatan kategori negatif yang diprediksi kategori negatif.

Tabel 1 Matriks konfusi untuk peubah respon dengan 2 kategori

Kategori Aktual	Kategori Prediksi	
	Positif	Negatif
Positif (Tidak Lulus)	*True Positive* (TP)	*False Negative* (FN)
Negatif (Lulus)	*False Positive* (FP)	*True Negative* (TN)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$Sensitivitas = \frac{TP}{TP+FN} \quad (15)$$

$$Spesifisitas = \frac{TN}{TN+FP}\,. \quad (16)$$

Ukuran keakuratan yang digunakan diantaranya akurasi, sensitivitas, dan spesifisitas yang diperoleh dari Tabel 1. Akurasi mengukur keefektifan klasifikasi secara keseluruhan dengan menghitung proporsi amatan yang diprediksi dengan tepat terhadap seluruh amatan. Sensitivitas mengukur akurasi klasifikasi amatan kategori positif sedangkan spesifisitas mengukur akurasi klasifikasi amatan kategori negatif. Penelitian ini fokus pada kategori positif sehingga selanjutnya akan ditekankan pada peningkatan nilai sensitivitas namun dengan tetap mempertimbangkan nilai akurasi dan spesifisitasnya. Selain ketiga ukuran keakuratan di atas, kurva ROC (*Receiver Operating Characteristic*) juga dapat digunakan untuk mengevaluasi kinerja metode klasifikasi. Namun pada penelitian ini kurva ROC akan digunakan untuk memperoleh nilai *cut off* yang dapat mengoptimumkan kinerja klasifikasi. Grafik ini memiliki sumbu-*x* berupa *TP rate* (sensitivitas) dan sumbu-*y* berupa *FP rate* (1-spesifisitas). Kurva tersebut terbentuk dari performa metode pada semua kemungkinan nilai *cut off* kemudian dihubungkan oleh sebuah kurva. *Cut off* merupakan nilai yang menunjukkan batas hasil klasifikasi.

METODE

Data

Data yang digunakan dalam penelitian ini merupakan data mahasiswa tahun ajaran 2011/2012 sampai tahun ajaran 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. Peubah respon yang diamati adalah keberhasilan studi mahasiswa di SPs-IPB berupa status kelulusan dengan dua kategori, yaitu lulus dan tidak lulus. Mahasiswa yang tidak lulus meliputi mahasiswa yang *drop out* dan mengundurkan diri. Sedangkan peubah penjelasnya ditunjukkan pada Tabel 2.

Tabel 2 Daftar peubah penjelas yang digunakan

Peubah Penjelas	Keterangan
Jenis Kelamin (X1)	Laki-laki
	Perempuan
Status Perkawinan (X2)	Belum Menikah
	Menikah
Status Penerimaan (X3)	Biasa
	Percobaan
	Fast Track
	PMDSU
Status Perguruan Tinggi (PT) Asal (X4)	PT Negeri
	PT Swasta
	PT Luar Negeri
Sumber Biaya Pendidikan (X5)	Beasiswa
	Sendiri
Pekerjaan (X6)	Instansi Negeri
	Instansi Swasta
	Luar Negeri
	Tidak Bekerja
Program Studi (X7)	Sains
	Sosial
Usia Masuk SPs-IPB (X8)	Numerik
IPK Asal S1 (X9)	Numerik

Metode Penelitian

Analisis dilakukan menggunkaan perangkat lunak R 3.4.3 dengan paket e1071, DMwR, caret, dan ROCR dengan langkah-langkah sebagai berikut.

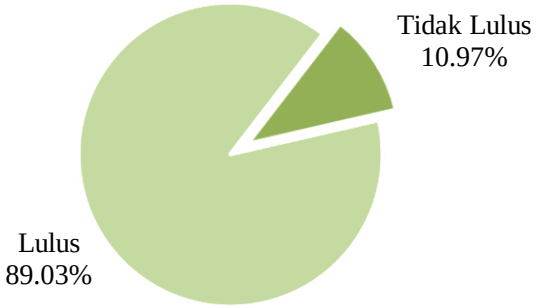
1. Melakukan eksplorasi untuk melihat karakteristik data keberhasilan studi mahasiswa program magister IPB.
2. Melakukan pembagian data uji dan data latih dengan persentase 80% data latih dan 20% data uji dengan perbandingan kelas minoritas dan mayoritas pada data latih dan data uji relatif sama dengan data asli.
3. Melakukan klasifikasi SVM pada data latih menggunakan SVM linier dan SVM non-linier, yaitu polinomial dan RBF serta mengevaluasi kinerja klasifikasi pada data uji menggunakan nilai akurasi, sensitivitas, dan spesifisitas.
4. Melakukan penanganan data tidak seimbang menggunakan SMOTE pada data latih dengan nilai k untuk k-tetangga terdekat yaitu 5.
5. Melakukan klasifikasi menggunakan 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 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 dilakukan penanganan, setelah dilakukan penanganan dengan SMOTE, dan setelah dilakukan SMOTE menggunakan *cut off* terbaik.

HASIL DAN PEMBAHASAN

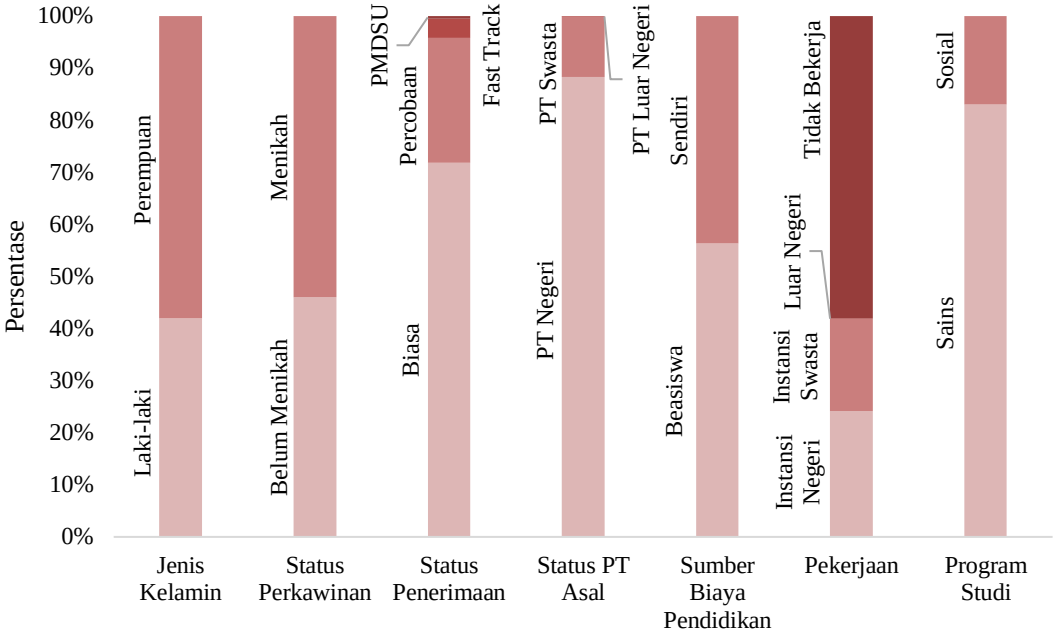
Karakteristik Mahasiswa Program Magister IPB

Data yang digunakan pada penelitian ini tidak mengikutsertakan data mahasiswa yang masih aktif berkuliah sehingga analisis hanya dilakukan pada data mahasiswa yang sudah memiliki status kelulusan baik lulus maupun tidak lulus. Keberhasilan studi mahasiswa diprediksi menggunakan status kelulusan tersebut sebagai peubah responnya. Berdasarkan Gambar 4, terdapat 4408 (89.03%) mahasiswa berstatus lulus dan 543 (10.97%) mahasiswa berstatus tidak lulus. Terlihat bahwa mahasiswa yang berstatus lulus memiliki persentase yang jauh lebih besar, sehingga menunjukkan adanya ketidakseimbangan kategori pada data. Pemeriksaan ada tidaknya ketidakseimbangan kategori pada data yang dianalisis dilakukan karena kondisi tersebut dapat mempengaruhi hasil klasifikasi.



Gambar 4 *Pie chart* banyaknya mahasiswa berdasarkan status kelulusan

Peubah penjelas yang digunakan pada penelitian ini yaitu berupa karakteristik dan latar belakang pendidikan mahasiswa yang terdiri dari 7 peubah bertipe kategorik dan 2 peubah bertipe numerik. Gambar 5 menunjukkan persentase mahasiswa pada setiap kategori peubah penjelas. Terlihat bahwa mahasiswa magister SPs-IPB didominasi oleh mahasiswa berjenis kelamin perempuan dengan persentase 58.07% sedangkan mahasiswa laki-laki memiliki persentase sebesar 41.93%. Selanjutnya untuk status perkawinan, terlihat bahwa persentase mahasiswa yang sudah menikah yaitu sebesar 53.99%. Persentase tersebut lebih besar dibandingkan persentase mahasiswa yang belum menikah yaitu 46.01%.



Gambar 5 *Stack column chart* banyaknya mahasiswa pada setiap kategori peubah penjelas

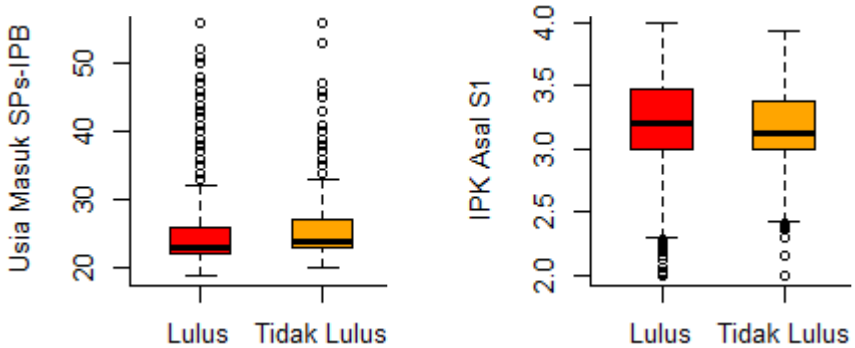
Mahasiswa baru di SPs-IPB dapat diterima dengan status biasa, percobaan, *fast track*, ataupun PMDSU. Status biasa diberikan ketika mahasiswa baru dapat memenuhi semua persyaratan dari SPs-IPB. Mahasiswa dengan status ini memiliki persentase terbesar yaitu 71.82%. Status percobaan diberikan kepada mahasiswa yang tidak memenuhi semua persyaratan namun memiliki prestasi tertentu. Jika pada akhir semester 1 mahasiswa dengan status percobaan tidak dapat memenuhi syarat IPK yang ditentukan, maka mahasiswa tersebut akan dikeluarkan dari SPs-IPB. Sebanyak 24.02% mahasiswa pada penelitian ini memiliki status percobaan. Status penerimaaan selanjutnya yaitu *fast track*. Mahasiswa yang diterima melalui jalur ini menjalani perkuliahan program magister bersamaan dengan perkuliahan program sarjana. Mahasiswa *fast track* memiliki persentase sebesar 3.66%. Status penerimaan terakhir yaitu Program Pendidikan Magister menuju Doktor untuk Sarjana Unggul (PMDSU). PMDSU merupakan program langsung S3 dengan melalui perkuliahan S2 selama 3 semester. Apabila mahasiswa tersebut dapat memenuhi persyaratan untuk langsung S3, maka tidak perlu menyelesaikan pendidikan S2. Hanya 0.50% mahasiswa yang diterima melalui jalur ini.

Mahasiswa yang diterima di SPs-IPB dapat berasal dari dalam maupun luar negeri. PT dalam negeri dibagi menjadi PT negeri dan PT swasta. Mahasiswa yang menempuh pendidikan sarjana di dalam negeri dengan status PT negeri memiliki persentase terbesar yaitu 88.31% sedangkan mahasiswa yang berasal dari PT swasta memiliki persentase sebesar 11.53%. PT luar negeri memiliki persentase terkecil yaitu 0.16%. Karakteristik mahasiswa selanjutnya juga dapat dilihat dari sumber pembiayaan pendidikan program magister. Sumber biaya pendidikan dapat dibagi menjadi dua yaitu dibiayai oleh suatu lembaga berupa beasiswa dan dibiayai sendiri oleh mahasiswa yang bersangkutan. Mahasiswa yang menerima beasiswa memiliki persentase 56.29%. Nilai tersebut lebih besar dibandingkan mahasiswa yang membiayai sendiri pendidikannya yaitu dengan persentase sebesar 43.71%. Jenis

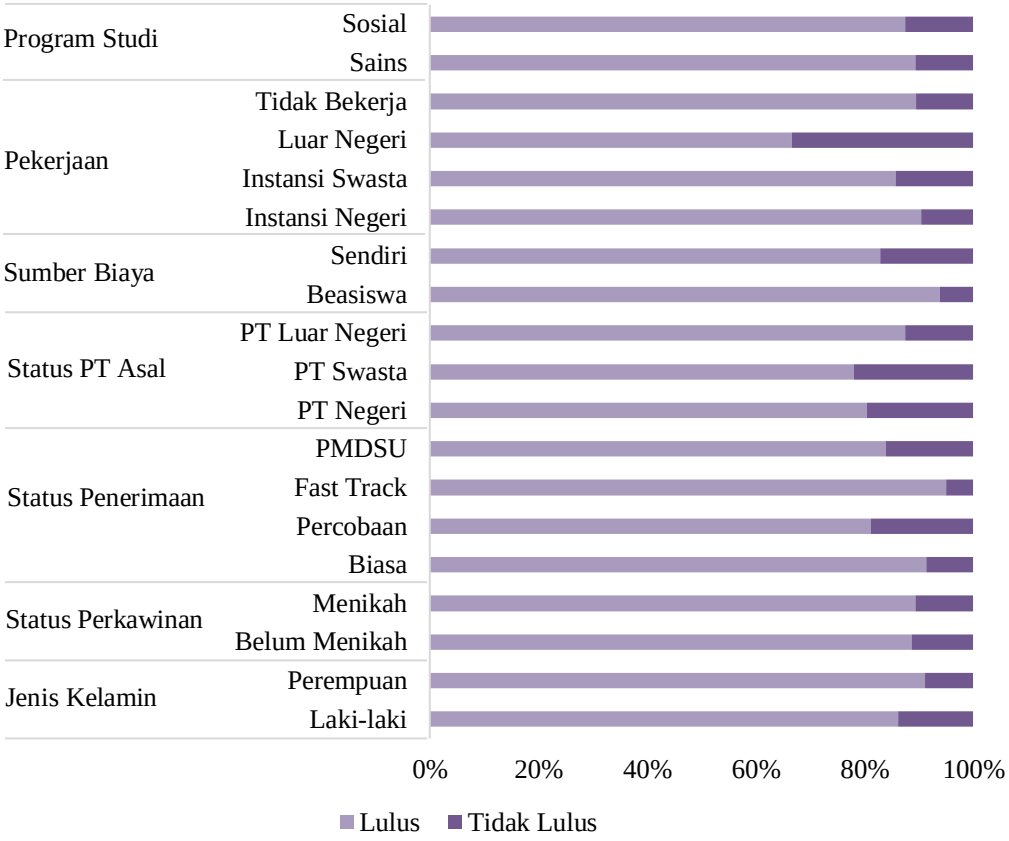
pekerjaan mahasiswa SPs-IPB dibagi menjadi 4 kategori, yaitu bekerja di instansi negeri, instansi swasta, instansi luar negeri, dan tidak bekerja. Mahasiswa yang tidak bekerja memiliki persentase terbesar yaitu 58.09% sedangkan persentase terendah dimiliki mahasiswa yang bekerja di instansi luar negeri yaitu 0.06%. Karakteristik terakhir yaitu program studi yang dipilih oleh mahasiswa. Program studi dikelompokkan menjadi 2 yaitu sains dan sosial. Mahasiswa yang memilih program studi sains memiliki persentase 83.01% sedangkan sisanya yaitu 16.99% memilih program studi sosial.

Peubah penjelas dengan tipe numerik yang digunakan pada penelitian ini adalah usia masuk SPs-IPB dan IPK asal S1. Gambar 6(a) menunjukkan *boxplot* untuk peubah usia. Terlihat bahwa kotak *boxplot* untuk mahasiswa yang lulus dan tidak lulus memiliki lebar yang sama dikarenakan memiliki nilai jangkauan antar kuartil yang sama yaitu 4 tahun. Nilai median usia masuk SPs-IPB untuk mahasiswa yang tidak lulus lebih tinggi dibandingkan mahasiswa yang lulus. Mahasiswa yang lulus dan tidak lulus memiliki nilai kuartil 3 berturut-turut sebesar 26 tahun dan 27 tahun. Hal tersebut menunjukkan bahwa pada masing-masing kategori 75% mahasiswa masuk SPs-IPB saat berusia kurang dari kedua nilai tersebut. Rentang usia masuk SPs-IPB pada kedua kategori status kelulusan juga tidak berbeda jauh. Mahasiswa dengan kategori lulus memiliki usia minimum 19 tahun dan mahasiswa dengan kategori tidak lulus memiliki usia minimum 20 tahun, sedangkan kedua kategori tersebut memiliki usia maksimum yang sama yaitu 56 tahun. Hal ini mengindikasikan bahwa usia tidak menjadi faktor pembeda pada kategori kelulusan mahasiswa magister SPs-IPB.

Gambar 6(b) menunjukkan *boxplot* untuk peubah IPK asal S1. Terlihat bahwa mahasiswa yang lulus memiliki kotak *boxplot* yang lebih lebar. Hal tersebut menunjukkan nilai IPK mahasiswa yang lulus lebih beragam. Kemudian nilai median IPK untuk kategori lulus lebih besar dibandingkan kategori tidak lulus. Rentang nilai IPK asal untuk mahasiswa yang lulus dan tidak lulus tidak jauh berbeda. Kedua kategori memiliki nilai IPK minimum yang sama yaitu 2.00 sedangkan IPK maksimum untuk mahasiswa yang lulus adalah 4.00 dan untuk 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 mahasiswa yang memiliki IPK asal yang rendah juga dapat berpotensi lulus.



Gambar 6 *Boxplot* untuk peubah penjelas (a) usia masuk SPs-IPB dan (b) IPK asal S1 berdasarkan status kelulusan



Gambar 7 *Stack bar chart* banyaknya mahasiswa pada setiap kategori peubah penjelas berdasarkan status kelulusan

Selajutnya akan dibahas persentase mahasiswa kategori tidak lulus setiap kategori peubah penjelas yang ditunjukkan pada Gambar 7. Karakteristik mahasiswa yang tidak lulus dapat dilihat dari besar persentase kategori pada setiap peubah penjelas. Terlihat bahwa mahasiswa berjenis kelamin laki-laki memiliki persentase tidak lulus yang jauh lebih besar yaitu 13.82% sedangkan mahasiswa berjenis kelamin perempuan yang tidak lulus memiliki persentase sebesar 8.90%. Dilihat dari status penerimaannya, Mahasiswa dengan status penerimaan percobaan memiliki persentase tidak lulus terbesar yaitu 18.84%. Kemudian diikuti dengan status penerimaan PMDSU, biasa, dan *fast track*. Lalu untuk status PT asal, status PT swasta memiliki persentase mahasiswa tidak lulus terbesar yaitu 21.89%. Selanjutnya berdasarkan sumber biaya pendidikan, terlihat bahwa mahasiswa yang membiayai pendidikannya sendiri memiliki persentase mahasiswa tidak lulus yang jauh lebih besar yaitu 17.10% sedangkan mahasiswa penerima beasiswa yang tidak lulus memiliki persentase sebesar 6.21%. Kemudian jika dilihat dari segi pekerjaan, mahasiswa yang bekerja di instansi luar negeri memiliki persentase tidak lulus yang paling besar yaitu 33.33%. Namun, karena mahasiswa yang bekerja di instansi luar negeri hanya terdiri dari 3 orang, maka mahasiswa yang bekerja di instansi swasta yang dianggap memiliki persentase tidak lulus terbesar. Persentase mahasiswa tidak lulus berdasarkan status perkawinan dan program studi tidak berbeda jauh antar kategorinya.

Kinerja Klasifikasi SVM

Sebelum dilakukan pemodelan SVM, data terlebih dahulu dibagi menjadi data latih dan data uji. Data latih memiliki komposisi 80% data awal dan data uji memiliki komposisi 20% data awal. Proporsi kategori mayoritas dan kategori minoritas pada data uji dan data latih disesuaikan dengan proporsi kategorinya pada data awal yaitu 90% untuk kategori mayoritas dan 10% untuk kategori minoritas. Hal ini diterapkan supaya persentase kategori mayoritas dan minoritas pada data latih dan data uji dapat merepresentasikan kondisi ketidakseimbangan kategori pada data awal. Komposisi pembagian data ditunjukkan pada Tabel 3. Selanjutnya dilakukan pemodelan pada data latih. Jenis SVM yang digunakan untuk pemodelan yaitu SVM linier dan SVM non-linier. Pemodelan dengan SVM non-linier menggunakan fungsi kernel polinomial dan RBF. Kemudian pembagian data latih dan data uji serta pemodelan dilakukan pengacakan sebanyak 100 kali untuk memeriksa kekonsistenan metode SVM dalam mengklasifikasikan data keberhasilan studi mahasiswa.

Tabel 3 Komposisi pembagian data latih dan data uji pada masing-masing kategori data

Kategori	Data Latih	Data Uji	Total
Total Data	3962 (80.00%)	989 (20.00%)	4951 (100.00%)
Minoritas (Tidak Lulus)	435 (10.98%)	108 (10.92%)	
Mayoritas (Lulus)	3527 (89.02%)	881 (89.08%)	

Pemodelan SVM dilakukan menggunakan nilai parameter *default* seperti yang ditunjukkan pada Tabel 4. Rataan hasil kinerja klasifikasi berbagai jenis SVM pada data uji menggunakan nilai akurasi, sensitivitas, dan spesifisitas ditunjukkan pada Tabel 5. Ketiga jenis SVM memiliki 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 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 dapat diklasifikasikan dengan tepat. Hal ini akan merugikan SPs-IPB jika menerima mahasiswa yang berpotensi besar tidak lulus. Oleh karena itu, selanjutnya akan dilakukan penanganan ketidakseimbangan kategori pada data untuk meningkatkan kemampuan SVM dalam mengklasifikasikan mahasiswa tidak lulus.

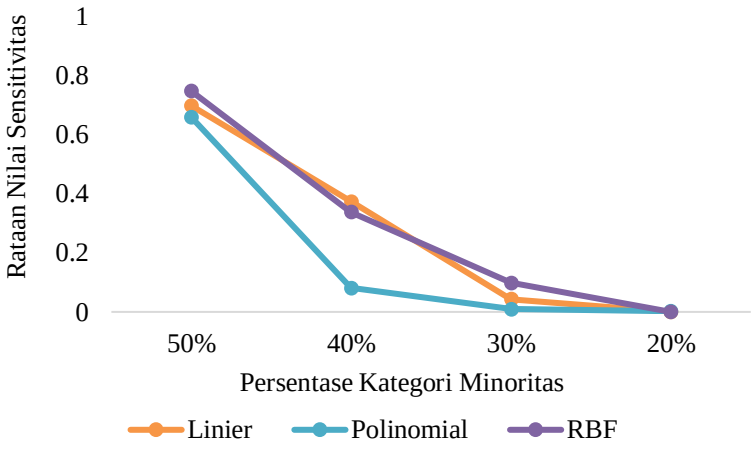
Tabel 4 Nilai parameter yang digunakan dalam melakukan pemodelan berbagai jenis SVM

Jenis SVM	Parameter		
	C	d	γ
Linier	1	-	-
Polinomial	1	3	-
RBF	1	-	0.0667

Tabel 5 Rataan kinerja klasifikasi pada data uji berbagai jenis SVM tanpa penanganan ketidakseimbangan kategori pada data

Jenis SVM	Kinerja Klasifikasi (%)		
	Akurasi	Sensitivitas	Spesifisitas
Linier	89.08	0.00	100.00
Polinomial	89.08	0.00	100.00
RBF	89.08	0.00	100.00

Ketika terdapat ketidakseimbangan kategori pada data, SVM akan cenderung mengklasifikasikan kategori minoritas ke dalam kategori mayoritas. Pada penelitian ini, mahasiswa yang berstatus tidak lulus akan diklasifikasikan lulus, sehingga sensitivitas yang dihasilkan bernilai 0.00%. Sebelumnya telah dilakukan simulasi untuk melihat kinerja klasifikasi dari ketiga jenis SVM pada berbagai persentase kategori minoritas. Pemodelan SVM dilakukan pada persentase kategori minoritas 50%, 40%, 30%, dan 20%. Pada Gambar 8 terlihat bahwa semakin kecil persentase kategori minoritas, rataan nilai sensitivitas juga semakin kecil. Saat persentase kategori minoritasnya 30%, terlihat rataan nilai sensitivitas untuk ketiga jenis SVM kurang dari 20%. Bahkan jenis SVM polinomial menghasilkan rataan nilai sensitivitas sebesar 0%. Oleh karena itu, dapat dikatakan kinerja klasifikasi SVM mulai mengalami penurunan ketika persentase kategori minoritasnya 30% sedangkan pada data keberhasilan studi ini persentase kategori minoritasnya adalah 10%. Pembagian data latih dan data uji memiliki perbandingan kategori minoritas dan mayoritas yang sama pada 100 ulangannya sehingga hasil kinerja klasifikasi memiliki nilai yang sama. Hal ini mengakibatkan penentuan jenis SVM yang tepat untuk mengklasfikasikan keberhasilan studi mahasiswa tidak dapat dilakukan karena hasil kinerja klasifikasi tidak dapat dibandingkan satu sama lain.



Gambar 8 Rataan nilai sensitivitas berbagai jenis SVM pada berbagai persentase kategori minoritas

Kinerja Klasifikasi SVM dengan SMOTE

Nilai sensitivitas yang sangat rendah pada hasil klasifikasi SVM pada bagian sebelumnya mengindikasikan perlunya dilakukan penanganan. Penelitian ini menerapkan metode SMOTE digunakan untuk mengatasi ketidakseimbangan kategori pada data keberhasilan studi mahasiswa SPs-IPB tersebut. Pada prinsipnya, metode SMOTE membangkitkan data buatan pada kategori minoritas dengan memanfaatkan konsep k-tetangga terdekat dengan nilai k yang akan digunakan yaitu 5. Data buatan dibangkitkan di antara amatan kategori minoritas dan salah satu tetangga terdekatnya yang dipilih secara acak sampai diperoleh perbandingan banyaknya kategori minoritas dan mayoritas yang relatif seimbang seperti yang ditunjukkan pada Tabel 6.

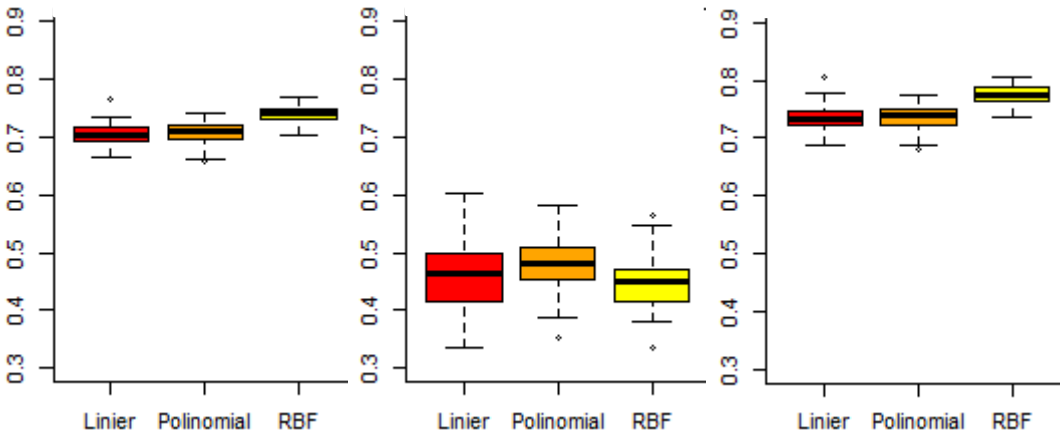
Tabel 6 Komposisi pembagian data latih sebelum dan sesudah melalui tahap SMOTE

Kategori	Data Latih		Data Latih Hasil SMOTE	
Minoritas	435	(10.98%)	3480	(49.67%)
Mayoritas	3527	(89.02%)	3527	(50.33%)
Total	3962	(100.00%)	7007	(100.00%)

Selanjutnya, SMOTE diterapkan pada 100 data latih yang sebelumnya sudah terbentuk untuk dievaluasi kembali dengan metode SVM. Tabel 7 menunjukkan rataan kinerja klasifikasi pada data uji berbagai jenis SVM setelah melalui tahap SMOTE. 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 kategori pada data dengan menggunakan SMOTE berhasil meningkatkan kemampuan SVM dalam mengklasifikasikan mahasiswa tidak lulus. Selanjutnya, jika sebelum melalui tahap SMOTE jenis SVM yang tepat untuk mengklasifikasikan keberhasilan studi mahasiswa tidak dapat ditentukan karena kinerja klasifikasi yang dihasilkan memiliki nilai yang sama, maka setelah melalui tahap SMOTE hal tersebut dapat dilakukan. Terlihat bahwa nilai rata-rata akurasi dan rata-rata spesifisitas tertinggi dihasilkan dari jenis SVM RBF yaitu berturut-turut sebesar 73.97% dan 77.54%. Namun nilai rata-rata sensitivitas tertinggi dihasilkan dari SVM polinomial yaitu sebesar 47.98%. Oleh karena itu, dapat dikatakan bahwa jenis SVM yang tepat digunakan untuk mengklasifikasikan keberhasilan studi mahasiswa adalah SVM RBF. Hal ini dikarenakan nilai akurasi hasil SVM RBF sebesar 73.97%, nilai sensitivitasnya mencapai 44.84% dan nilai spesifisitasnya mencapai 77.54%.

Tabel 7 Rataan kinerja klasifikasi pada data uji 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



(a) (b) (c)

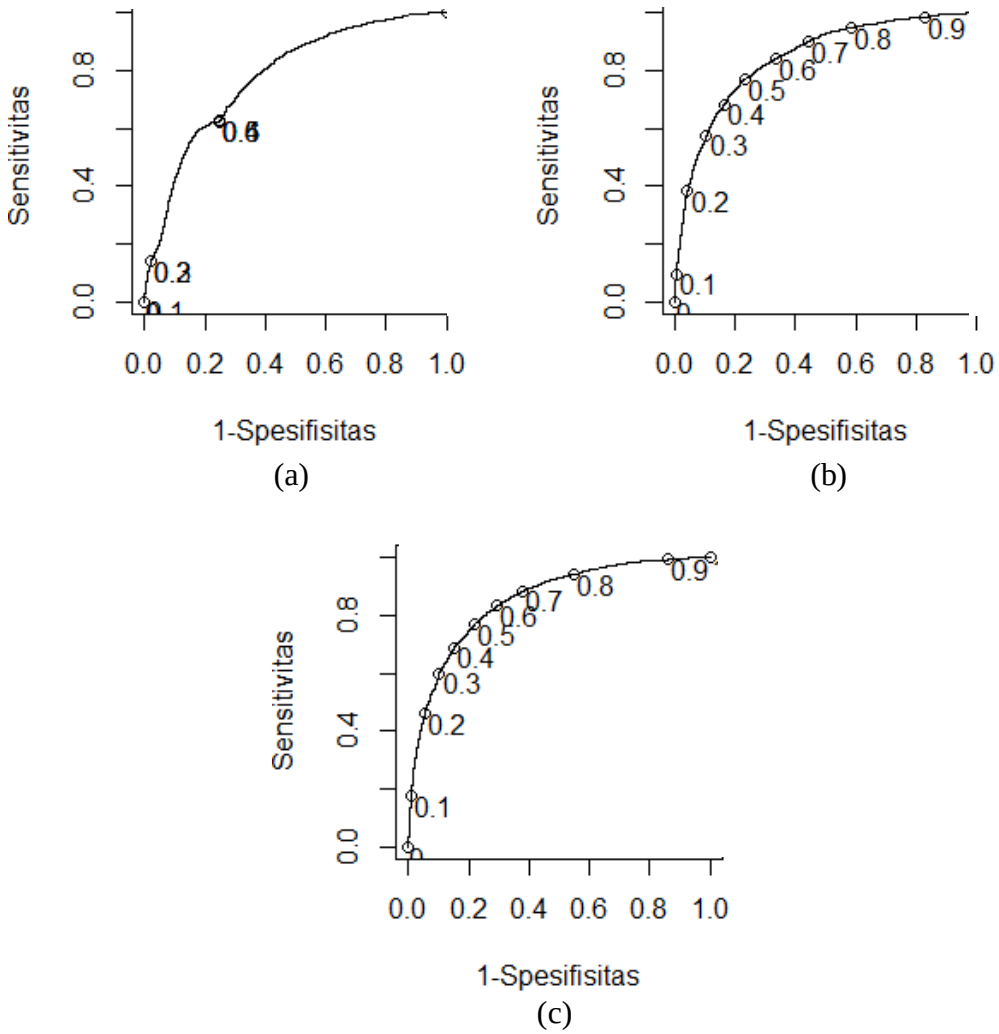
Gambar 9 *Boxplot* nilai (a) akurasi, (b) sensitivitas, dan (c) spesifisitas berbagai jenis SVM setelah melalui tahap SMOTE

Gambar 9 menunjukkan *boxplot* kinerja klasifikasi SVM setelah melalui tahap SMOTE untuk ketiga jenis SVM. Pada Gambar 9(a) yang menggambarkan nilai akurasinya terlihat bahwa *boxplot* untuk ketiga jenis SVM tidak terlalu lebar serta tidak mengandung banyak pencilan. Hal ini menandakan nilai akurasi memiliki keragaman yang kecil. Hal ini juga terjadi pada nilai spesifisitasnya yang ditunjukkan pada Gambar 9(c). Berbeda halnya dengan nilai sensitivitas seperti yang ditunjukkan pada Gambar 9(b). Lebar *boxplot* untuk ketiga jenis SVM berbeda. SVM linier memiliki *boxplot* paling lebar yang berarti memiliki keragaman paling besar. Sedangkan SVM polinomial dan RBF memiliki lebar kotak *boxplot* yang hampir sama. Secara keseluruhan terlihat bahwa *boxplot* sensitivitas lebih lebar dibandingkan dengan *boxplot* akurasi dan spesifisitas. Hal ini menunjukkan bahwa nilai sensitivitas yang dihasilkan oleh SVM memiliki keragaman yang lebih besar dibandingkan dengan keragaman nilai akurasi dan spesifisitasnya.

Kinerja Klasifikasi SVM dengan SMOTE menggunakan *Cut Off* Terbaik

Setelah dilakukan penanganan ketidakseimbangan kategori pada data menggunakan SMOTE nilai sensitivitas yang dihasilkan sudah mengalami peningkatan namun dapat dikatakan masih rendah. Nilai tersebut dapat ditingkatkan lagi dengan mengubah nilai *cut off* yang digunakan. Nilai *cut off* tersebut diperoleh melalui kurva ROC. Penentuan nilai *cut off* tidak menggunakan data latih melainkan data keseluruhan. Gambar 10 menunjukkan kurva ROC untuk SVM linier, polinomial, dan RBF beserta nilai *cut off* untuk masing-masing jenis SVM. Sebuah metode klasifikasi akan memiliki kinerja klasifikasi yang baik ketika kurva ROC yang terbentuk lebih condong ke arah pojok kiri atas. Nilai *cut off* yang terletak di pojok kiri atas akan menghasilkan nilai sensitivitas yang tinggi dan nilai 1-spesifisitas yang rendah, sehingga penentuan nilai *cut off* didasarkan pada ketentuan tersebut. Terlihat pada Gambar 10(a) yaitu kurva ROC untuk SVM linier,

nilai *cut off* yang letaknya di pojok kiri atas memiliki nilai 0.4, 0.5, dan 0.6. Kemudian untuk SVM polinomial, nilai *cut off* yang digunakan yaitu 0.4, 0.5, 0.6, dan 0.7, seperti yang terlihat pada Gambar10(b). Lalu untuk SVM RBF, seperti yang ditunjukkan pada Gambar 10(c), nilai *cut off* yang digunakan adalah 0.3, 0.4, 0.5, 0.6, dan 0.7. 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 10 Kurva ROC untuk jenis SVM (a) linier, (b) polinomial, dan (c) RBF beserta nilai *cut off* untuk masing-masing jenis SVM

Hasil rataan kinerja klasifikasi SVM pada berbagai nilai *cut off* ditunjukkan pada Tabel 8. 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* terbaik untuk setiap jenis SVM didasarkan pada ketiga ukuran keakuratan tersebut. Nilai sensitivitas diharapkan dapat meningkat dibandingkan ketika menggunakan nilai *cut off default* yaitu 0.5 namun tanpa mengorbankan penurunan nilai akurasi dan spesifisitas yang terlalu besar.

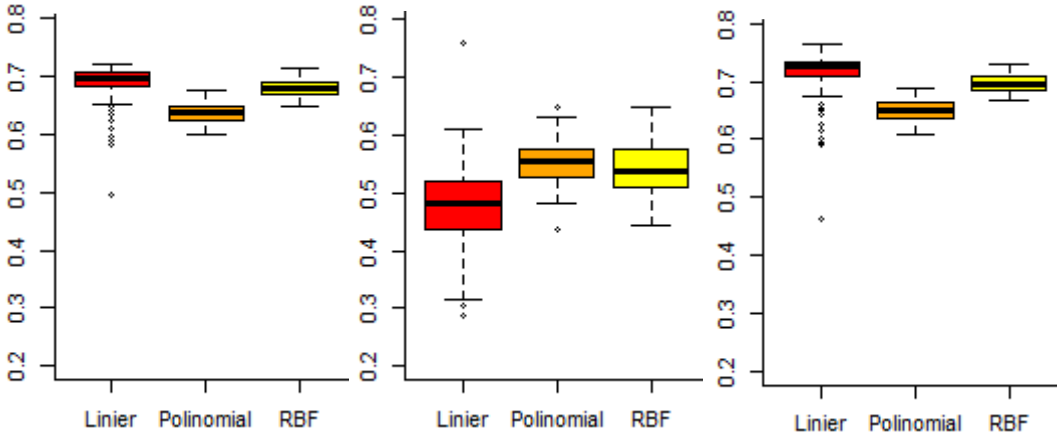
SVM linier memiliki rataan kinerja klasifikasi paling baik pada *cut off* 0.6 karena rataan nilai sensitivitasnya meningkat yang semula bernilai 45.09% menjadi 47.26%. Selain itu, penurunan rataan nilai akurasi dan spesifisitasnya pun hanya sekitar 2%. Kemudian untuk SVM polinomial, rataan kinerja klasifikasi paling baik dihasilkan pada nilai *cut off* yang sama dengan SVM linier, yaitu 0.6. Terlihat bahwa rataan nilai sensitivitasnya mengalami peningkatan, yang semula 46.09% menjadi 55.24%. Ketika menggunakan *cut off* 0.7, rataan nilai sensitivitasnya adalah 63.44%. Nilai ini lebih besar dibandingkan ketika menggunakan *cut off* 0.6. Namun, rataan nilai akurasi dan spesifisitasnya terlalu rendah yaitu kurang dari 60%, sehingga *cut off* yang dipilih tetap 0.6. Jenis SVM selanjutnya yaitu SVM RBF. Berdasarkan lima nilai *cut off* yang dicoba, rataan kinerja klasifikasi terbaik juga dihasilkan pada *cut off* 0.6. Terlihat bahwa rataan nilai sensitivitas mengalami peningkatan dari 46.06% menjadi 54.14%. Ketika menggunakan *cut off* 0.7, rataan nilai sensitivitasnya adalah 63.53%. Nilai ini lebih besar dibandingkan ketika menggunakan *cut off* 0.6. Namun, rataan nilai akurasi dan spesifisitasnya terlalu rendah yaitu sekitar 60%, sehingga *cut off* yang dipilih tetap 0.6. Selanjutnya hasil terbaik dari setiap jenis SVM dibandingkan seperti yang ditunjukkan pada Tabel 10.

Tabel 8 Rataan kinerja klasifikasi berbagai jenis SVM pada data uji dengan 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 9 Rataan kinerja klasifikasi pada data uji 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



Gambar 11 *Boxplot* nilai (a) akurasi, (b) sensitivitas, dan (c) spesifisitas berbagai jenis SVM setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik

Tabel 9 menunjukkan hasil rataan kinerja klasifikasi SVM pada data uji setelah melalui tahap SMOTE dan menggunakan *cut off* terbaik. Terlihat bahwa nilai rata-rata akurasi dan rata-rata spesifisitas tertinggi dihasilkan dari jenis SVM linier yaitu berturut-turut sebesar 68.80% dan 71.44%. Namun nilai rata-rata sensitivitas tertinggi dihasilkan dari SVM polinomial yaitu sebesar 55.24%. Sehingga dapat dikatakan bahwa SVM linier memiliki kinerja yang lebih baik. Tetapi berdasarkan Gambar 11, terlihat bahwa *boxplot* nilai akurasi dan spesifisitas SVM linier mengandung banyak pencilan 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 dipililh sebagai jenis SVM yang tepat digunakan untuk mengklasifikasikan keberhasilan studi mahasiswa. 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 pada penelitian ini.

Tabel 10 menunjukkan rataan nilai sensitivitas untuk ketiga jenis SVM ketika pemodelan dilakukan tanpa penanganan, penanganan menggunakan SMOTE, dan penanganan menggunakan SMOTE serta *cut off* terbaik. 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.

Tabel 10 Perbandingan rataan nilai sensitivitas pada data uji berbagai jenis SVM pada 3 tahapan analisis

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

Model akhir yang digunakan untuk memprediksi keberhasilan studi mahasiswa SPs-IPB diperoleh dari pemodelan SVM RBF dengan *cut off* 0.6 setelah melalui tahap SMOTE. Kinerja klasifikasi yang dihasilkan ditunjukkan pada Tabel 11. Nilai akurasi menunjukkan bahwa kemampuan model akhir dalam mengklasifikasikan mahasiswa secara tepat adalah sebesar 76.51%. Nilai sensitivitas menunjukkan bahwa mahasiswa yang berstatus tidak lulus yang dapat diklasifikasikan dengan tepat sebesar 83.08%. Sedangkan nilai spesifisitas menunjukkan bahwa mahasiswa berstatus lulus yang dapat diklasifikasikan dengan tepat yaitu sebesar 70.03%.

Tabel 11 Kinerja klasifikasi SVM RBF dengan *cut off* 0.6 pada data keseluruhan setelah melalui tahap SMOTE

Kinerja Klasifikasi	Nilai (%)
Akurasi	76.51
Sensitivitas	83.08
Spesifisitas	70.03

SIMPULAN

Data keberhasilan studi mahasiswa magister SPs-IPB yang digunakan pada penelitian ini merupakan data yang berkategori tidak seimbang karena banyaknya mahasiswa dengan kategori tidak lulus (10.97%) jauh lebih sedikit dibandingkan mahasiswa dengan kategori lulus (89.03%). Pemodelan SVM pada data dengan kategori tidak seimbang tersebut memberikan kinerja klasifikasi yang kurang memuaskan ditunjukkan dengan nilai sensitivitas yang sangat rendah. Oleh karena itu, dilakukan penanganan untuk mengatasi kondisi ketidakseimbangan tersebut dengan menggunakan SMOTE. Hasilnya menunjukkan bahwa SMOTE berhasil meningkatkan nilai sensitivitas hingga mencapai lebih dari 40%. Upaya lain yang dilakukan untuk meningkatkan nilai sensitivitas yaitu dengan menggunakan nilai *cut off* terbaik dari setiap jenis SVM dalam melakukan pemodelan. Berdasarkan evaluasi yang dilakukan, model terbaik yang digunakan untuk memprediksi keberhasilan studi mahasiswa SPs-IPB diperoleh dari pemodelan SVM RBF dengan *cut off* 0.6 setelah melalui tahap SMOTE dengan nilai sensitivitas 54.14%.

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, Djemaa HK, Alitouche TA. 2012. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*. 2(10):17-28.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2001. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16:211-257.

Ganganwar V. 2012. An overview of claassification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*. 2:42-47.

Han H, Wang WY, Mao BH. 2005. Borderline-SMOTE: a new over-sampling method in imbalance data sets learning. *Springer-Verlag Berlin Heidelberg*. 878-887.

James G, Witten D, Hastie T, Tibshirani R. 2012. *An Introduction to Statistical Learning with Applications in R*. New York (US): Springer.

[IPB] Institut Pertanian Bogor. 2013. *Katalog Sekolah Pascasarjana 2013*. Bogor(ID): IPB.

Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets : a review. *GESTS International Transactions on Computer Science and Engineering*. (20).

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

Raykov T, Marcoulides GA. 2008. *An Introduction to Applied Multivariate Analysis*. New York (US): Routledge.

Schölkopf B, Smola AJ. 2000. *Learning with Kernels.* Massachusetts (US): MIT Press.

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS *et al*. 2008. Top 10 algorithms in data mining. *Knowledge Information System.* 14:1-37.

RIWAYAT HIDUP

Penulis dilahirkan di Jakarta pada 19 Oktober 1996 dan merupakan putri kedua dari dua bersaudara, dari pasangan Bapak Yusron jamal dan Ibu Chafidah. Penulis menempuh pendidikan sekolah menengah atas di Madrasah Aliyah Negeri 4 Jakarta dan lulus pada tahun 2014. Pada tahun yang sama penulis diterima di Institut Pertanian Bogor melalui jalur Seleksi Bersama Masuk Perguruan Tinggi (SBMPTN) pada departemen Statistika dengan minor Ilmu Ekonomi. Selama menjadi mahasiswa departemen statistika, penulis aktif mengikuti organisasi dan kepanitiaan. Penulis pernah menjadi Ketua Departemen Kesekretariatan Himpunan Mahasiswa Profesi Gamma Sigma Beta periode 2016/2017. Selain itu kepanitiaan yang pernah diikuti oleh penulis diantaranya anggota divisi acara Pekan Olahraga Statistika tahun 2015, sekretaris divisi kesekretariatan *The 11st* Statistika Ria tahun 2016, anggota divisi medis *Welcome Ceremony of Statistics* tahun 2016, anggota divisi kesekretariatan *The 4th IPB Business Festival* tahun 2017, dan sekretaris divisi kesekretariatan Pesta Sains Nasional sub Kompetisi Statistika Junior tahun 2017. Pada tahun 2017 penulis melakukan praktik lapang di Pusat Data dan Sistem informasi Pertanian, Kementerian Pertanian.