Xplore (ISSN:2302-5751), Vol 2 No 1 (2018):49-60

Penerapan Metode *Resampling* Dan *K - Nearest Neighbor* Dalam Memprediksi Keberhasilan Studi Mahasiswa Program Magister IPB

Devi Andrian*, Agus M Soleh*, Hari Wijayanto*
*Departemen Statistika Institut Pertanian Bogor

Abstrak*—Sekolah Pascasarjana IPB dipercaya menghasilkan lulusan yang berkualitas dan berdaya saing tinggi, namun berdasarkan data yang ada, terdapat sebagian kecil mahasiswa yang tidak lulus, baik mengundurkan diri maupun *Drop Out*. Hal tersebut perlu ditangani salah satunya dengan menerapkan metode klasifikasi *K - Nearest Neighbor* (KNN) untuk memprediksi potensi keberhasilan studi mahasiswa. Peubah respon yang digunakan adalah status keberhasilan studi mahasiswa, yaitu lulus dan tidak lulus. Sedangkan peubah penjelas adalah profil dan latar belakang pendidikan S1 mahasiswa. Terdapat ketakseimbangan pada data yang diperoleh, kelas tidak lulus berjumlah jauh lebih sedikit dibandingkan kelas lulus. Hal ini dapat menurunkan nilai akurasi klasifikasi pada kelas minoritas (sensitivitas) sehingga dilakukan penanganan ketakseimbangan data dengan menggunakan *Random Over Sampling* (ROS), *Random Under Sampling* (RUS), dan *Random Over - Under Sampling* (ROUS). Nilai evaluasi hasil klasifikasi KNN (k = 1 hingga 6), mengalami peningkatan nilai sensitivitas setelah disertai penanganan ketakseimbangan data, meskipun nilai akurasi dan spesifisitas mengalami penurunan. Metode KNN dengan k = 6 disertai ROS merupakan skema klasifikasi KNN terbaik dalam memprediksi potensi keberhasilan studi mahasiswa program magister IPB dibandingkan skema lainnya. Nilai rataan dan median sensitivitas sebesar 0.89 dan 0.89, nilai rataan dan median spesifisitas sebesar 0.75 dan 0.75, serta nilai rataan dan median akurasi sebesar 0.77 dan 0.77.

Kata kunci*—Magister IPB; KNN; *Resampling

I. PENDAHULUAN

A. Latar Belakang

Sekolah Pascasarjana Institut Pertanian Bogor (SPs IPB) secara terstruktur dimulai pada tahun 1975. SPs IPB pada awalnya hanya terdiri dari 7 program studi dan lebih menekankan pada program magister sains. Seiring dengan perkembangan sumber daya dan mutu pendidikan yang ada, pada tahun 1978 program doktor resmi dibuka dan saat ini SPs IPB menyelenggarakan 65 program studi magister dan 43 program studi doktor. Selain itu, banyak perguruan tinggi dan lembaga nasional maupun internasional yang bekerjasama dengan SPs - IPB, serta banyak lulusan yang telah bekerja di dalam maupun luar negeri. Hal ini dilandasi oleh rasa kepercayaan terhadap lulusan SPs - IPB yang dianggap memiliki daya saing yang tinggi. Sebagai upaya untuk mempertahankan reputasi tersebut, salah satunya dapat dilakukan dengan cara melakukan klasifikasi untuk mengetahui potensi keberhasilan studi mahasiswa program magister saat menempuh proses pendidikannya sehingga mahasiswa yang berpotensi mengalami kegagalan studi dapat diberikan suatu tindakan pencegahan terjadinya kegagalan studi dimasa mendatang oleh pihak terkait.

Mahasiswa program magister yang berpotensi lulus dan tidak lulus dapat diprediksi dengan analisis klasifikasi menggunakan data profil mahasiswa program magister. Analisis klasifikasi merupakan salah satu bagian dari data mining yang bertujuan untuk memprediksi label kategori benda yang tidak diketahui sebelumnya, dalam membedakan antara objek yang satu dengan yang lainnya berdasarkan atribut atau fitur (Ngai et al. (2011)). Salah satu analisis klasifikasi yang dapat digunakan adalah analisis klasifikasi *K - Nearest Neighbor* (KNN). KNN sendiri merupakan metode klasifikasi yang paling dasar dan sederhana (Parvin, Alizadeh, and Bidgoli (Parvin et al.)). Meskipun sederhana, KNN dianggap sebagai salah satu dari sepuluh algoritma klasifikasi data mining yang terbaik (Wu et al. (2008)), bahkan dalam beberapa kasus, Kuramochi and Karypis (2005) mengemukakan bahwa KNN mengungguli performa *Support Vector Machine* (SVM) yang memiliki skema klasifikasi yang lebih

canggih.

Berdasarkan data keberhasilan studi yang diperoleh dari SPs-IPB, diketahui bahwa terdapat sebagian kecil mahasiswa program magister yang tidak berhasil lulus. Hal tersebut mengindikasikan bahwa terdapat ketakseimbangan data antara mahasiswa yang lulus (mayoritas) dan tidak lulus (minoritas). Ketakseimbangan ini akan berdampak pada hasil prediksi klasifikasi, karena hampir semua analisis klasifikasi menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas dibandingkan kelas minoritas saat ada ketakseimbangan data (Gu et al. (2016)).

Metode *resampling* merupakan salah satu metode yang dapat digunakan dalam menangani adanya ketakseimbangan data. Siringoringo (2017) dalam penelitiannya menyatakan bahwa integrasi metode *resampling* pada KNN efektif dalam menangani ketakseimbangan kelas dan meningkatkan nilai akurasi, sensitivitas, dan spesifisitas pada hasil klasifikasi.

Penelitian ini akan menerapkan metode integrasi *resampling* dan KNN pada prediksi keberhasilan studi mahasiswa program magister IPB. Selain itu akan dilakukan tahapan persiapan data (*preprocessing*) pada data untuk meningkatkan performa dan penyesuaian data masukan pada analisis klasifikasi yang digunakan.

B. Tujuan

Tujuan dari penelitian ini adalah memprediksi potensi keberhasilan studi mahasiswa program magister IPB tahun 2011 hingga 2015 menggunakan skema metode KNN dengan dan tanpa *resampling*, serta menentukan skema klasifikasi KNN terbaik.

II. TINJAUAN PUSTAKA

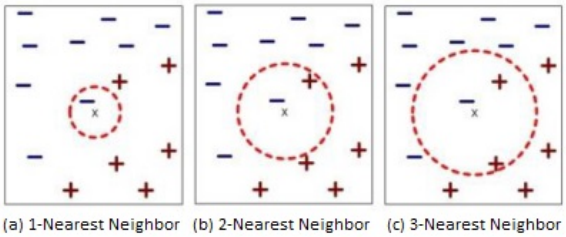
A. Resampling

Secara umum, terdapat tiga pendekatan untuk menangani dataset tidak seimbang, yaitu pendekatan pada level data, level algoritma, dan menggabungkan atau memasangkan (*ensemble*) metode (Yap, Rani, Rahman, Fong, Khairudin, and Abdullah (Yap et al.)). Pendekatan pada level data dapat berupa penerapan berbagai metode *resampling*, maupun metode sintesis pada data untuk memperbaiki perbedaan distribusi kelas pada data latih (Saifudin and Wahono (2015)). Metode *resampling* yang dapat digunakan adalah *Random Over Sampling* (ROS), *Random Under Sampling* (RUS), dan gabungan keduanya, *Random Over - Under Sampling* (ROUS).

Metode RUS melakukan perhitungan selisih antara banyak anggota kelas mayoritas dan minoritas data kelas mayoritas akan dihapus secara acak, hingga jumlah kelas mayoritas sama dengan minoritas (Saifudin and Wahono (2015)). Sedangkan metode ROS melakukan peningkatan ukuran kelas minoritas dengan cara mensintesis sampel baru atau langsung mereplikasi secara acak dataset latih (Yu et al. (2013)). Pada metode RUS kemungkinan adanya data penting yang terbuang menjadi lebih besar dan pada metode ROS kemungkinan akan memperberat proses komputasi karena penambahan jumlah data (Garcia, Sanchez, and Mollineda (Garcia et al.)).

B. K - Nearest Neighbor

K - Nearest Neighbor (KNN) merupakan salah satu metode klasifikasi yang menggunakan prinsip ketetanggaan dalam memprediksi kelas data baru (Siringoringo (2018)). Kelas data baru ditentukan berdasarkan kelas tetangga terdekat yang paling banyak muncul (Karno (Karno)), namun apabila terdapat beberapa k tetangga terdekat yang memiliki frekuensi kemunculan yang sama, maka akan dipilih satu tetangga terdekat secara acak, seperti diilustrasikan pada Gambar 1 dan ditunjukkan pada persamaan (1).



Gambar 1. Ilustrasi *K Nearest Neighbor*

$$y' = argmaxv \sum_{(x_i,y_i)\in D_z} I(v = y_i) \qquad (1)$$

dimana :

y' = label kelas data uji

y_i = label kelas data latih ke - i

x_i = peubah bebas data latih ke i

D_z = himpunan K tetangga terdekat pada data latih terhadap data uji

$I()$ = fungsi indikator yang meberikan nilai 1 jika benar dan 0 jika salah

Proses klasifikasi pada KNN dipengaruhi oleh tiga faktor, yaitu banyaknya data latih, metode perhitungan jarak dan banyaknya k tetangga terdekat yang digunakan (Wu et al. (2008)). Selain itu Houben et al. (1997) yang menyatakan bahwa KNN sangat dipengaruhi oleh jenis perhitungan jarak yang digunakan dan bobot pada setiap data latih. Banyaknya K pada KNN akan berdampak pada hasil klasifikasi yang diperoleh, nilai k yang terlalu besar akan menyebabkan hasil klasifikasi menjadi sulit diperoleh atau kabur, sedangkan saat nilai k terlalu kecil (k=1), hasilnya akan terasa kaku (Indrayanti et al. (2017)) sehingga penentuan k yang tepat perlu dilakukan agar diperoleh hasil klasifikasi yang akurat.

Selain penentuan nilai k tetangga terdekat, penentuan metode perhitungan jarak yang digunakan pada KNN juga akan berpengaruh terhadap hasil klasifikasi. Jenis data yang dimiliki akan berpengaruh terhadap penentuan metode jarak yang sebaiknya digunakan. Perhitungan jarak antar observasi berupa data campuran kuantitatif dan kualitatif dapat dilakukan dengan menggunakan koefisien kemiripan umum Gower.

Koefisien kemiripan Gower merupakan koefisien kemiripan yang dikemukakan oleh Gower (1971) dalam penelitiannya yang berjudul *A General Coefficient of Similarity and Some of Its Properties*. Koefisien kemiripan Gower dapat digunakan untuk melihat kemiripan antar observasi dengan melakukan perhitungan jarak pada setiap peubah acak yang ada sesuai dengan skala pengukuran perubah acak tersebut. Secara umum, persamaan kemiripan Gower ditunjukkan pada persamaan (2).

$$s(x_i, x_j) = \frac{\sum_{k=1}^{p} s_k(x_{ik}, x_{jk})\delta(x_{ik}, x_{jk})w_k}{\sum_{k=1}^{p} \delta(x_{ik}, x_{jk})w_k} \quad (2)$$

$$\delta(x_{ik}, x_{jk}) = \begin{cases} 1; x_{ik}, x_{jk} \in R \\ 0; lainnya \end{cases} \quad (3)$$

dimana :

$s(x_i, x_j)$ = koefisien kemiripan Gower antara observasi ke i dan j

$s_k(x_{ik}, x_{jk})$ = koefisien kemiripan antara observasi ke i dan j pada peubah k

$\delta(x_{ik}, x_{jk})$ = kemungkinan perbandingan peubah k observasi ke i dan j

w_k = bobot pilihan yang menyatakan kepentingan variabel. w_k = 1

x_{ik}, x_{jk} = nilai observasi ke i dan j pada peubah ke - k

Koefisien kemiripan s_k pada setiap peubah dihitung berdasarkan skala pengukuran peubah k. Perhitungan nilai s_k pada skala nominal, ordinal, interval, dan rasio secara berurutan ditunjukkan pada persamaan (4), (5), dan (6).

$$Nominal; s_k(x_{ik}, x_{jk}) = \begin{cases} 1; x_{ik} = x_{jk} \\ 0; x_{ik} \neq x_{jk} \end{cases} \quad (4)$$

$$Ordinal; s_k(x_{ik}, x_{jk}) = 1 - \frac{|r(x_{ik}) - r(x_{jk})|}{R(r(x_{mk}))} \quad (5)$$

$$Numerik; s_k(x_{ik}, x_{jk}) = 1 - \frac{|x_{ik} - x_{jk}|}{R(x_{mk})} \quad (6)$$

dimana :

$r(x_{mk})$ = rank dari nilai observasi ke m, peubah ordinal k

$\max_m x_{mk}$ = rank maksimum dari seluruh nilai peubah ordinal k

$\min_m x_{mk}$ = rank minimum dari seluruh nilai peubah ordinal k

$R(r(x_{mk}))$ = jarak dari rank maksimum dan minimum peubah ordinal k

$R(x_{mk})$ = jarak dari maksimum dan minimum peubah ordinal k

Penentuan tetangga terdekat KNN diperoleh berdasarkan nilai jarak ketakmiripan antar observasi sehingga nilai koefisien kemiripan Gower perlu ditransformasi menjadi nilai koefisien ketakmiripan dengan menggunakan persamaan (7).

$$d(x_{ik}, x_{jk}) = 1 - s(x_{ik}, x_{jk}) \quad (7)$$

dimana : $d(x_{ik}, x_{jk})$ = koefisien ketakmiripan Gower observasi ke i dan j peubah k

C. Validasi Hasil Klasifikasi

Mengevaluasi hasil klasifikasi pada kasus data tidak seimbang, akurasi akan lebih ditekankan pada kelas minoritas sehingga validasi dapat dilakukan dengan menghitung nilai TPrate/Recall/Sensitivitas untuk mengukur akurasi kelas minoritas (Irawan (2015)). Nilai evaluasi terhadap hasil klasifikasi dirangkum dalam suatu tabel matriks konfusi, seperti diilustrasikan pada Tabel 1.

Tabel I
MATRIKS KONFUSI

Kelas Aktual	Kelas Prediksi	
	Positif (Kelas = 0)	Negatif (Kelas = 1)
Positif (Kelas = 0)	TP	FN
Negatif (Kelas = 1)	FP	TN

True Positive $> \alpha_2 1$ (TP) dan *True Negative* (TN) merupakan banyaknya amatan kelas positif dan negatif yang diklasifikasikan sesuai dengan kelas sebenarnya. *False Negative* (FN) dan *False Positive* (FP) merupakan banyaknya amatan kelas positif dan negatif yang diklasifikasikan berbeda dengan kelas sebenarnya. Berdasarkan nilai TP, TN, FP, dan FN, dapat diperoleh nilai akurasi, sensitivitas (TPrate), dan spesivisitas (TNrate). Perhitungan nilai evaluasi klasifikasi ditunjukkan pada persamaan berikut.

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

$$Sensitivitas = \frac{TP}{TP + FN} \quad (9)$$

$$Spesifisitas = \frac{TN}{TN + FP} \quad (10)$$

III. METODOLOGI

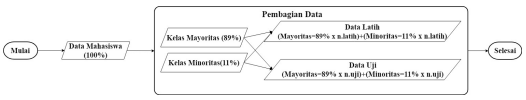
A. Data

Data yang digunakan dalam penelitian ini adalah data sekunder mengenai profil akademik mahasiswa program magister tahun ajaran 2011/2012 hingga 2015/2016 yang diperoleh dari SPs IPB. Data awal yang diperoleh sebanyak 6116 amatan, namun setelah dilakukan proses *cleaning data*, maka jumlah amatan yang diperoleh sebanyak 4951 amatan dengan 1 peubah respon dan 9 peubah penjelas. Peubah respon yang diamati adalah status kelulusan mahasiswa program magister dengan banyak mahasiswa lulus (Y = 1) dan tidak lulus (Y = 0) masing-masing sebanyak 4408 dan 543 mahasiswa. Mahasiswa yang berstatus tidak lulus mencakup mahasiswa yang mengundurkan diri dan *Drop Out* (DO). Sedangkan peubah penjelas yang digunakan terdiri dari peubah kategorik, yaitu Jenis Kelamin (X1), Status Pernikahan (X2), Jalur Penerimaan (X3), Status Perguruan Tinggi S1 (X4), Sumber Biaya Pendidikan S2 (X5), Kelompok Pekerjaan (X6), dan Kelompok Program Studi S2 (X7), serta peubah numerik, yaitu Usia Diterima S2 (X8) dan IPK S1 (X9).

B. Prosedur Analisis Data

Proses analisis data dilakukan dengan menggunakan perangkat lunak R 3.4.3. Tahapan analisis data dilakukan sebagai berikut :

1) Melakukan proses persiapan data sebelum dianalisis (*preprocessing*) dengan melakukan proses *Data Cleaning* / *Data Filtering* pada data yang dimiliki untuk mendeteksi, memperbaiki, menghapus atau bahkan menambahkan nilai dengan tujuan untuk meminimalisir ketidak konsistenan pada dataset yang tersedia.
2) Membagi data menjadi data latih dan data uji dan dilakukan sebanyak 100 ulangan. Proporsi data latih sebesar 80% dan data uji sebesar 20%. Skema pembagian data dijelaskan pada Gambar 2.

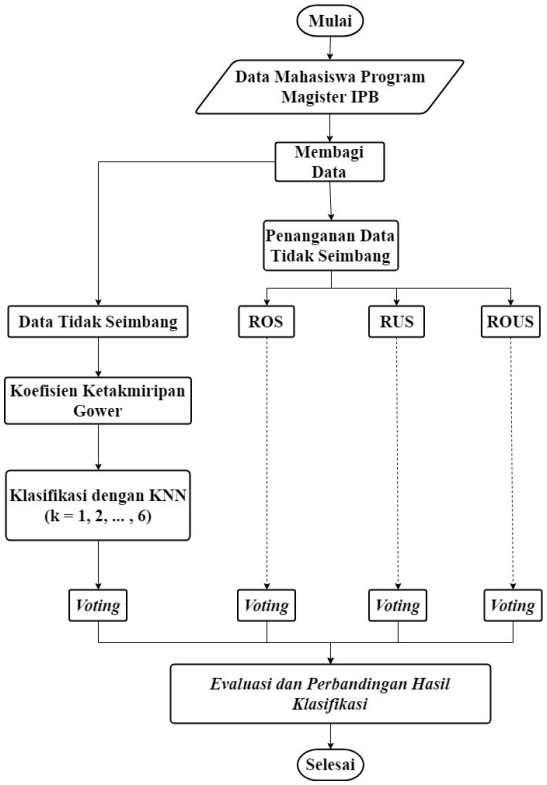


Gambar 2. Skema Pembagian Data

3) Melakukan penanganan data tak seimbang dengan metode *resampling* (ROS, RUS, dan ROUS), menggunakan *package* ROSE sehingga banyaknya data kelas lulus dan tidak lulus menjadi seimbang dengan perbandingan 50 : 50.

4) Melakukan analisis klasifikasi KNN, menggunakan *package* KODAMA.
   a) Menghitung nilai koefisien ketakmiripan Gower antara data latih dan data uji menggunakan persamaan (2) hingga (7).
   b) Menentukan k tetangga terdekat untuk setiap data uji.
   c) Melakukan *voting* menggunakan persamaan (1). Kelas dengan nilai *voting* tertinggi (mayoritas) akan dipilih sebagai kelas bagi data uji.
5) Mengevaluasi dan membandingkan hasil setiap integrasi metode *Resampling* dan KNN menggunakan matriks konfusi.

Diagram alir integrasi metode Resampling + KNN ditunjukkan pada Gambar 3.

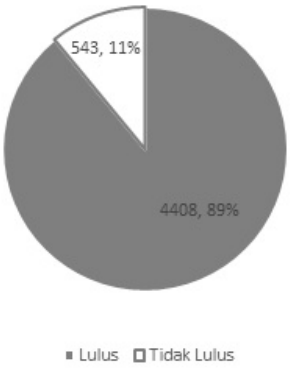


Gambar 3. Diagram Alir Resampling + KNN

IV. HASIL DAN PEMBAHASAN

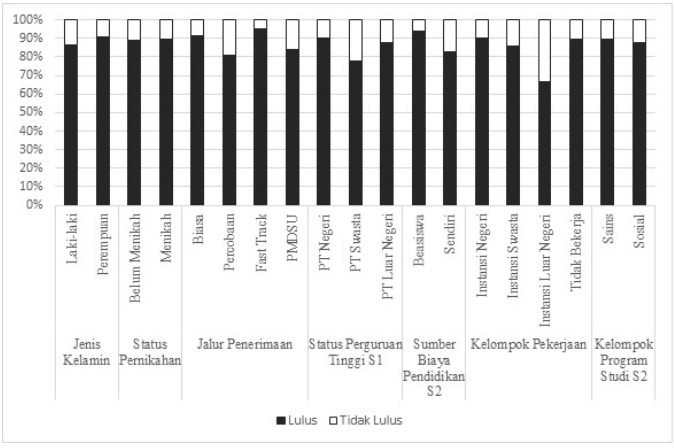
A. Analisis Deskriptif

Data awal profil akademik mahasiswa program magister tahun ajaran 2011/2012 hingga 2015/2016 yang diperoleh dari SPs IPB sebanyak 6116 amatan, akan tetapi setelah dilakukan proses cleaning data, jumlah amatan yang diperoleh menurun menjadi 4951 amatan yang secara umum teridiri dari 4408 mahasiswa berstatus lulus dan 543 mahasiswa berstatus tidak lulus. Secara eksploratif, gambaran status kelulusan mahasiswa program magister ditunjukkan pada Gambar 4.



Gambar 4. *Piechart* Status Kelulusan Mahasiswa

Berdasarkan Gambar 4, dapat diketahui bahwa persentase mahasiswa berstatus lulus sebesar 89% dan mahasiswa berstatus tidak lulus sebesar 11%. Berdasarkan hal tersebut dapat diketahui bahwa mahasiswa berstatus lulus jauh lebih banyak dibandingkan mahasiswa tidak lulus. Hal ini mengindikasikan bahwa adanya ketidak seimbangan kelas respon pada data profil akademik mahasiswa program magister IPB.



Gambar 5. Presentase Kelulusan Mahasiswa Tiap Peubah Penjelas Kategorik

Presentase status kelulusan mahasiswa proram magister IPB untuk setiap peubah kategorik dapat dilihat pada Gambar 5. Berdasarkan Gambar 5 dapat diketahui bahwa pada peubah jenis kelamin, persentase mahasiswa yang tidak lulus lebih besar terjadi pada kelas laki - laki dengan nilai 13.82%, dibandingkan kelas perempuan yang bernilai 8.90%. Namun, keduanya memiliki nilai persentase tidak lulus yang relatif rendah, yakni bernilai dibawah 15%. Selanjutnya berdasarkan peubah status pernikahan, dapat diketahui bahwa mahasiwa yang belum menikah memiliki persentase tidak lulus sebesar 11.33%, nilai ini lebih besar dibandingkan mahasiswa yang sudah menikah dengan persentase tidak lulus sebesar 10.66%. Lalu berdasarkan peubah jalur penerimaan, mahasiswa dengan jalur percobaan memiliki persentase tidak lulus paling tinggi dibandingkan dengan mahasiswa yang masuk melalui jalur penerimaan lainnya (biasa, fast track, dan PMDSU) dengan nilai 18.84% dan mahasiswa dengan jalur penerimaan fast track memiliki nilai persentase tidak lulus paling kecil dengan nilai 4.97%.

Berdasarkan peubah status perguruan tinggi S1, persentase tidak lulus tertinggi hingga terendah berturut turut terjadi pada kelas perguruan tinggi swasta, perguruan tinggi luar negeri, dan perguruan tinggi negeri dengan nilai 21.89%, 12.50%, dan 9.54%. Lalu berdasarkan peubah sumber biaya pendidikan S2, persentase tidak lulus yang terjadi pada mahasiswa penerima beasiswa bernilai 6.21%, nilai ini bernilai lebih kecil dibandingkan persentase tidak lulus pada mahasiswa yang membiayai pendidikan S2nya dengan biaya pribadi dengan nilai sebesar 17.10%. Kemudian, berdasarkan peubah kelompok pekerjaan, mahasiswa yang bekerja di instansi luar negeri memiliki persentase tidak lulus yang jauh lebih besar dibanding kelas lainnya (instansi negeri, swasta, dan tidak bekerja) dengan nilai sebesar 33.33%, berbeda dengan persentase tidak lulus pada kelas mahasiswa yang bekerja pada instansi negeri yang memiliki presentasi tidak lulus terkecil dengan nilai 9.55%. Berdasarkan peubah kelompok program studi S2, mahasiswa yang kuliah pada kelompok sosial memiliki persentase tidak lulus sebesar 12.49% dan mahasiswa yang kuliah pada kelompok sains memiliki nilai yang lebih kecil, yaitu sebesar 10.66%. Secara umum, terlihat bahwa persentase mahasiswa tidak lulus pada setiap perubah penjelas kategorik memiliki persentase yang jauh lebih kecil dibandingkan persentase mahasiswa yang lulus. Hal ini selaras dengan informasi gambaran umum status kelulusan mahasiswa magister IPB yang diperoleh dari Gambar 4 sebelumnya.

Tabel II
DESKRIPSI PEUBAH PENJELAS NUMERIK

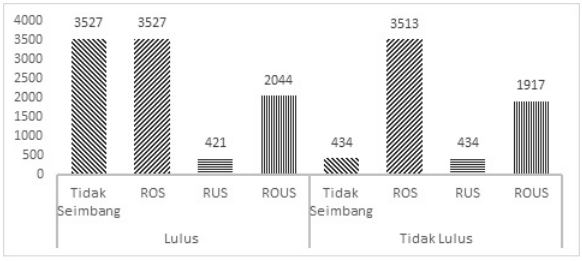
Peubah	Min.	Q1	Median	Rata-rata	Q3	Maks.
Umur	19	22	23	25	27	56
- L	19	22	23	25	26	56
- TL	20	23	24	26	27	56
IPK	2.00	3.00	3.20	3.20	3.48	4.00
- L	2.00	3.00	3.21	3.21	3.48	4.00
- TL	2.00	3.00	3.13	3.14	3.38	3.94

Peubah penjelas numerik yang digunakan pada penelitian ini berupa umur masuk S2 dan Indeks Prestasi Kumulatif (IPK) S1. Deskripsi mengenai kedua peubah tersebut dapat dilihat pada Tabel 2. Berdasarkan informasi pada Tabel 3, dapat diketahui bahwa secara umum umur mahasiswa yang diterima saat masuk program magister IPB rata rata berumur 25 tahun dan berada pada rentang 19 - 56 tahun. Lalu mahasiswa yang berstatus lulus dan tidak lulus masing - masing memiliki rata rata umur 25 dan 26 tahun saat diterima program magister, dengan rentang umur 19 - 56 tahun dan 20 - 56 tahun. Selain itu, secara umum rata rata IPK S1 mahasiswa saat diterima program magister sebesar 3.20 dan berada pada rentang nilai 2.00 - 4.00. Mahasiswa yang berstatus lulus dan tidak lulus program magister IPB masing masing memiliki nilai rata rata IPK saat diterima program magister IPB sebesar 3.21 dan 3.14, dengan rentang nilai masing- masing sebesar 2.00 4.00 dan 2.00 3.94.

B. Data Latih Dan Data Uji

Pembagian data latih dan data uji dilakukan secara acak pada data yang tersedia, sebanyak seratus kali ulangan. Pembagian data pada setiap ulangan menghasilkan data latih sebanyak 3961 amatan dan data uji sebanyak 990 amatan. Namun, berdasarkan besarnya perbedaan amatan antar kelas pada peubah status kelulusan yang mengindikasikan adanya ketakseimbangan antara kelas lulus dan tidak lulus,

maka perlu dilakukan upaya penyeimbangan data untuk meminimalisir kesalahan klasifikasi karena didominasi oleh hasil klasifikasi kelas mayoritas. Sehingga dilakukan upaya penyeimbangan data dengan menerapkan tiga jenis metode resampling pada data latih, menggunakan metode ROS, RUS, dan ROUS, dengan rincian seperti ditunjukkan pada Gambar 5.



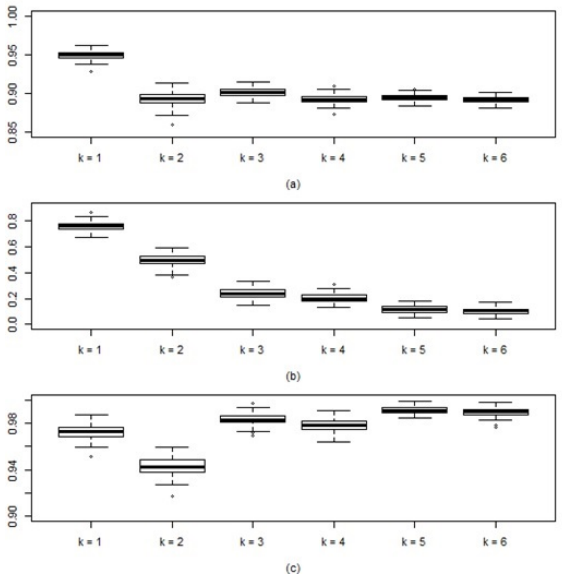
Gambar 6. Data Latih Sebelum dan Setelah *Resampling*

Berdasarkan diagram batang pada Gambar 5, dapat diketahui bahwa terdapat perubahan banyaknya amatan pada kelas minoritas maupun kelas mayoritas saat dilakukan penyeimbangan data dengan metode ROS, RUS dan ROES. Metode ROS melakukan penanganan ketidak seimbangan data dengan menambah amatan pada kelas tidak lulus menjadi sebanyak 3513 amatan, mendekati banyaknya amatan kelas lulus. Lalu metode RUS melakukan penanganan ketidak seimbangan data dengan mengurangi amatan pada kelas lulus secara acak, hingga menjadi 421 amatan, mendekati banyaknya amatan kelas tidak lulus. Metode ROUS melakukan penanganan ketidak seimbangan data dengan mengurangi jumlah amatan kelas lulus menjadi 2044 amatan dan menambah jumlah amatan kelas tidak lulus menjadi 1917 amatan.

C. K - Nearest Neighbor

Proses klasifikasi data uji menggunakan metode KNN dilakukan pada seluruh data latih, baik disertai dengan penerapan resampling maupun tidak. Penerapan KNN dengan data latih tanpa menerapkan resampling, menghasilkan hasil klasifikasi seperti ditunjukkan pada Gambar 6.

Berdasarkan Gambar 6, dapat diketahui bahwa seiring bertambahnya nilai parameter k, nilai rataan sensitivitas cenderung mengalami penurunan pada nilai rataan dan ragamnya. Hal ini terlihat dari penurunan posisi dan penyempitan lebar pada diagram



Gambar 7. Diagram kotak garis hasil klasifikasi metode KNN tanpa resampling: (a) akurasi, (b) sensitivitas, dan (c) spesifisitas

kotak garis sensitivitas saat nilai k bertambah. Nilai sensitivitas terbaik diperoleh saat k = 1, dengan nilai rataan dan median sebesar 0.76. Sedangkan nilai spesifisitas memiliki nilai rataan yang relatif konstan seiring bertambahnya nilai k dan mengalami penyempitan lebar diagram kotak garis, yang mengindikasikan bahwa setiap bertambahnya nilai k, keragaman nilai spesifisitas cenderung menurun. Nilai spesifisitas terbaik diperoleh saat k = 5, dengan nilai rataan dan median sebesar 0.99. Lalu nilai akurasi memiliki nilai rataan terbesar saat k = 1 dengan nilai rataan sebesar 0.95 dan nilai rataan cenderung relatif konstan saat k = 2 dan seterusnya dengan nilai rataan sebesar 0.90. Selain itu, nilai keragaman akurasi relatif konstan saat nilai k bertambah. Sehingga dipilih nilai k = 1 sebagai nilai k terbaik dalam memprediksi keberhasilan studi mahasiswa program magister IPB pada data latih tanpa proses resampling karena memiliki nilai rataan sensitivitas terbesar dibanding k lainnya, degan rincian seperti ditunjukkan pada Tabel 3.

Penerapan KNN pada data latih dengan metode penyeimbangan ROS, menghasilkan hasil klasifikasi seperti ditunjukkan pada Gambar 7. Berdasarkan Gambar 7 diperoleh informasi bahwa nilai rataan sensitivitas cenderung meningkat dan nilai keragaman sensitivitas cenderung menurun saat bertambahnya nilai k. Hal tersebut terlihat dari naiknya

Tabel III
DESKRIPSI NILAI AKURASI, SENSITIVITAS, DAN SPESIFISITAS KNN TANPA RESAMPLING, K = 1

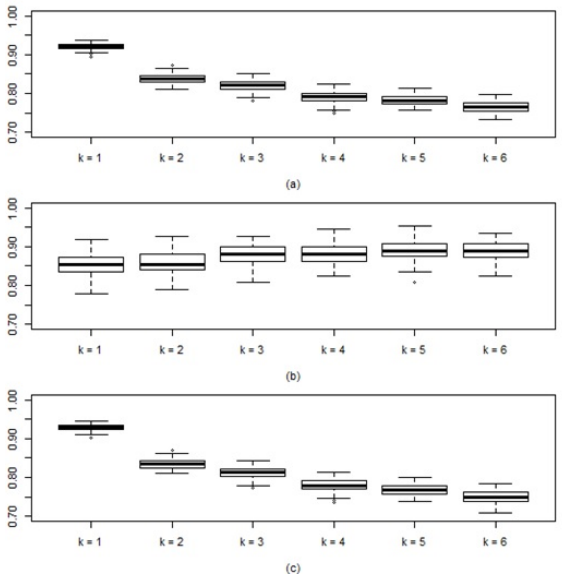
Evaluasi	Min.	Q1	Median	Rataan	Q3	Maks.
Akurasi	0.93	0.95	0.95	0.95	0.95	0.96
Sensitivitas	0.67	0.73	0.76	0.76	0.78	0.86
Spesifisitas	0.95	0.97	0.97	0.97	0.98	0.99

posisi dan penyempitan lebar diagram kotak garis sensitivitas saat nilai k bertambah. Nilai sensitivitas terbaik diperoleh saat nilai k = 6, dengan nilai rataan dan median sebesar 0.89. Namun nilai spesifisitas cenderung menurun saat bertambahnya nilai k, dengan nilai keragaman yang relatif konstan. Nilai spesifisitas terbaik diperoleh saat k = 1, dengan nilai rataan dan median sebesar 0.93. Begitu pun dengan nilai akurasi yang cenderung menurun saat bertambahnya nilai k, dengan nilai keragaman yang berfluktuatif. Nilai akurasi terbaik diperoleh saat menggunakan k = 1, dengan nilai rataan dan median sebesar 0.92. Sehingga dipilih nilai k = 6 sebagai nilai k terbaik dalam memprediksi keberhasilan studi mahasiswa program magister IPB pada data latih dengan metode penyeimbangan ROS karena memiliki nilai rataan sensitivitas terbesar dibanding k lainnya, degan rincian seperti ditunjukkan pada Tabel 4.

Tabel IV
DESKRIPSI NILAI AKURASI, SENSITIVITAS, DAN SPESIFISITAS KNN DENGAN ROS, K = 6

Evaluasi	Min.	Q1	Median	Rataan	Q3	Maks.
Akurasi	0.73	0.75	0.77	0.77	0.78	0.80
Sensitivitas	0.83	0.87	0.89	0.89	0.91	0.94
Spesifisitas	0.71	0.74	0.75	0.75	0.76	0.78

Penerapan KNN pada data latih dengan metode penyeimbangan RUS, menghasilkan hasil klasifikasi seperti ditunjukkan pada Gambar 8. Berdasarkan Gambar 8 diperoleh informasi bahwa nilai rataan sensitivitas cenderung menurun, namun nilai keragaman cenderung meningkat saat bertambahnya nilai k. Hal tersebut terlihat dari menurunnya posisi dan peningkatan lebar diagram kotak garis sensitivitas saat nilai k bertambah. Nilai sensitivitas

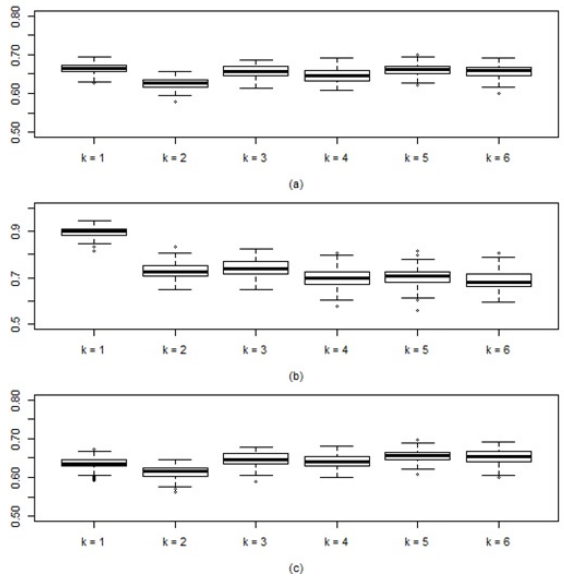


Gambar 8. Diagram Kotak Garis Hasil Klasifikasi Metode KNN dan ROS: (a) akurasi, (b) sensitivitas, dan (c) spesifisitas

terbaik diperoleh saat nilai k = 1, dengan nilai rataan sebesar 0.89 dan median sebesar 0.90. Sedangkan nilai spesifisitas memiliki nilai rataan yang berfluktuatif dan nilai keragaman yang relatif konstan saat bertambahnya nilai k. Nilai spesifisitas terbaik diperoleh saat nilai k = 5, dengan nilai rataan dan median sebesar 0.66. Lalu nilai akurasi memiliki perubahan nilai yang berfluktuatif, dengan nilai keragaman yang cenderung membesar saat bertambahnya nilai k. Nilai akurasi terbaik diperoleh saat nilai k = 1, 3, 5, dan 6 dengan nilai rataan sebesar 0.66 dan median sebesar 0.67. Sehingga dipilih nilai k = 1 sebagai nilai k terbaik dalam memprediksi keberhasilan studi mahasiswa program magister IPB pada data latih dengan metode penyeimbangan RUS karena memiliki nilai rataan sensitivitas terbesar dibanding k lainnya, dengan rincian seperti ditunjukkan pada Tabel 5.

Tabel V
DESKRIPSI NILAI AKURASI, SENSITIVITAS, DAN SPESIFISITAS KNN DENGAN RUS, K = 1

Evaluasi	Min.	Q1	Median	Rataan	Q3	Maks.
Akurasi	0.63	0.66	0.67	0.66	0.67	0.69
Sensitivitas	0.82	0.88	0.90	0.89	0.91	0.95
Spesifisitas	0.59	0.63	0.64	0.64	0.65	0.67

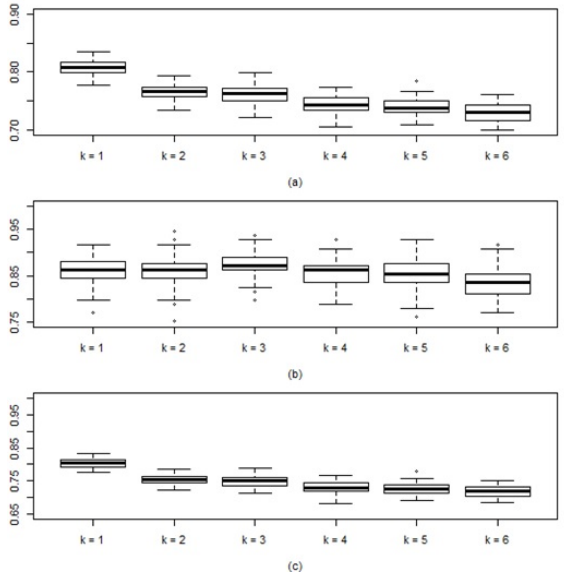


Gambar 9. Diagram Kotak Garis Hasil Klasifikasi Metode KNN dan RUS: (a) akurasi, (b) sensitivitas, dan (c) spesifisitas

Penerapan KNN pada data latih dengan metode penyeimbangan ROUS, menghasilkan hasil klasifikasi seperti ditunjukkan pada Gambar 9. Berdasarkan Gambar 9 diperoleh informasi bahwa nilai rataan sensitivitas cenderung meningkat saat k = 1 hingga k = 3, lalu mengalami penurunan saat k lebih dari 3. Selain itu nilai keragaman sensitivitas berubah secara fluktuatif saat nilai k bertambah. Nilai sensitivitas terbaik diperoleh saat nilai k = 3, dengan nilai rataan dan median sebesar 0.87. Sedangkan nilai rataan spesifisitas cenderung mengalami penurunan saat nilai k meningkat, dengan nilai keragaman yang relatif konstan. Hal tersebut terlihat dari penurunan posisi, serta lebar yang relatif konstan pada diagram kotak garis spesifisitas saat nilai k bertambah. Nilai spesifisitas terbaik diperoleh saat nilai k =1, dengan nilai rataan dan median sebesar 0.80. Begitu pun dengan nilai rataan dan keragaman akurasi yang cenderung menurun saat nilai k bertambah. Nilai akurasi terbaik diperoleh saatn nilai k = 1, dengan nilai rataan dan median sebesar 0.81. Sehingga dipilih nilai k = 3 sebagai nilai k terbaik dalam memprediksi keberhasilan studi mahasiswa program magister IPB pada data latih dengan metode penyeimbangan ROUS karena memiliki rataan nilai sensitivitas terbesar dibanding k lainnya, degan rincian seperti ditunjukkan pada Tabel 6.

Tabel VI
DESKRIPSI NILAI AKURASI, SENSITIVITAS, DAN SPESIFISITAS KNN DENGAN ROUS, K = 3

Evaluasi	Min.	Q1	Median	Rataan	Q3	Maks.
Akurasi	0.72	0.75	0.76	0.76	0.77	0.80
Sensitivitas	0.80	0.86	0.87	0.87	0.89	0.94
Spesifisitas	0.71	0.74	0.75	0.75	0.76	0.79



Gambar 10. Diagram Kotak Garis Hasil Klasifikasi Metode KNN dan ROUS: (a) akurasi, (b) sensitivitas, dan (c) spesifisitas

Berdasarkan hasil klasifikasi keberhasilan studi mahasiswa program magister IPB yang diperoleh dari metode KNN, baik dengan menggunakan resampling maupun tidak, dapat diketahui bahwa nilai sensitivitas terbaik diperoleh saat melakukan klasifikasi dengan metode integrasi KNN dengan k = 6, pada data latih dengan metode penyeimbangan ROS. Hal ini diperoleh berdasarkan nilai rataan dan median yang cukup tinggi dibandingkan hasil klasifikasi lainnya, yaitu sebesar 0.89 dan 0.89 (Tabel 8). Meskipun nilai rataan, median, seta lebar dan posisi diagram kotak garis sensitivitas KNN dengan k = 6 dan ROS tidak berbeda nyata dibandingkan KNN dengan k = 1 dan RUS, namun nilai rataan, median, dan selang nilai akurasi dan spesifisitas KNN dengan k = 6 dan ROS memiliki nilai yang cukup berbeda nyata dengan selisih nilai

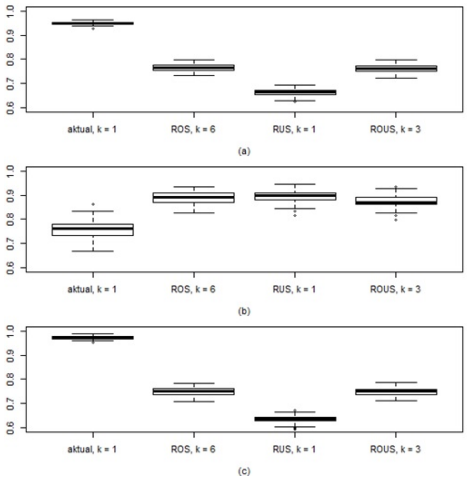
sekitar 0.10. Selain itu, posisi diagram kotak garis akurasi dan spesifisitas KNN dengan k = 6 dan ROS memiliki posisi yang lebih tinggi dibandingkan KNN dengan k = 1 dan RUS (Gambar 10). Hal ini mengindikasikan bahwa nilai akurasi dan spesifisitas KNN dengan k = 6 dan ROS bernilai lebih besar dibandingkan metode KNN dengan k = 1 dan RUS, namun metode penyeimbangan ROS membuat proses klasifikasi KNN menggunakan waktu komputasi yang lebih banyak dibandingkan tanpa menggunakan proses penyeimbangan maupun dengan proses penyeimbangan resampling lainnya, karena perhitungan jarak data uji pada setiap data latih menjadi lebih banyak dilakukan.

Tabel VII
PERBANDINGAN NILAI AKURASI, SENSITIVITAS, DAN SPESIFISITAS KNN PADA DATA LATIH DENGAN METODE RESAMPLING DAN TANPA RESAMPLING

Data Latih	Min.	Q1	Median	Rata-rata	Q3	Maks.
			Akurasi			
- Aktual, k = 1	0.93	0.95	0.95	0.95	0.95	0.96
- ROS, k = 6	0.73	0.75	0.77	0.77	0.78	0.80
- RUS, k = 1	0.63	0.66	0.67	0.66	0.67	0.69
- ROUS, k = 3	0.72	0.75	0.76	0.76	0.77	0.80
			Sensitivitas			
- Aktual, k = 1	0.67	0.73	0.76	0.76	0.78	0.86
- ROS, k = 6	0.83	0.87	0.89	0.89	0.91	0.94
- RUS, k = 1	0.82	0.88	0.90	0.89	0.91	0.95
- ROUS, k = 3	0.80	0.86	0.87	0.87	0.89	0.94
			Spesifisitas			
- Aktual, k = 1	0.95	0.97	0.97	0.97	0.98	0.99
- ROS, k = 6	0.71	0.74	0.75	0.75	0.76	0.78
- RUS, k = 1	0.59	0.63	0.64	0.64	0.65	0.67
- ROUS, k = 3	0.71	0.74	0.75	0.75	0.76	0.79

Lalu nilai spesifisitas terbaik diperoleh saat melakukan klasifikasi dengan metode KNN dengan k = 1, pada data latih tanpa menggunakan metode penyeimbangan. Hal ini diperoleh berdasarkan nilai rataan dan median yang lebih besar dibandingkan hasil klasifikasi lainnya, yaitu sebesar 0.97 (Tabel 8). Selain itu, nilai spesifisitas saat menggunakan KNN dengan k = 1 tanpa metode penyeimbangan memiliki lebar diagram kotak garis yang jauh lebih kecil dibandingkan dengan lebar diagram kotak garis spesifisitas lainnya (Gambar 10). Hal ini mengindikasikan bahwa nilai keragaman spesifisitas saat menggunakan KNN dengan k = 1 tanpa metode penyeimbangan bernilai jauh lebih kecil dibandingkan metode KNN dengan menggunakan metode penyeimbangan *resampling*.

Selanjutnya, untuk nilai akurasi terbaik diperoleh saat melakukan klasifikasi dengan metode KNN dengan k = 1, pada data latih tanpa menggunakan metode penyeimbangan. Hal ini diperoleh berdasarkan nilai rataan dan median yang bernilai jauh lebih besar dibandingkan hasil klasifikasi lainnya, yaitu sebesar 0.95 (Tabel 8). Selain itu, nilai akurasi saat menggunakan KNN dengan k = 1 tanpa metode penyeimbangan memiliki lebar diagram kotak garis yang jauh lebih kecil jika dibandingkan dengan lebar diagram kotak garis akurasi lainnya (Gambar 10). Hal ini mengindikasikan bahwa nilai keragaman akurasi saat menggunakan KNN dengan k = 1 tanpa metode penyeimbangan bernilai jauh lebih kecil dibandingkan metode KNN dengan menggunakan metode penyeimbangan *resampling*.



Gambar 11. Diagram Kotak Garis Hasil Klasifikasi Metode KNN Pada Data Latih dengan Metode *Resampling* dan Tanpa *Resampling*: (a) akurasi, (b) sensitivitas, dan (c) spesifisitas

V. SIMPULAN DAN SARAN

A. Simpulan

Pendugaan potensi keberhasilan studi mahasiswa program magister IPB dengan menggunakan metode KNN tanpa melakukan penyeimbangan data pada data latih dengan menggunakan nilai k = 1, menghasilkan nilai sensitivitas terbaik dengan nilai rataan

dan median sensitivitas sebesar 0.76. Penerapan metode penyeimbangan data dengan ROS, RUS dan ROUS sebelum proses klasifikasi dengan KNN menghasilkan nilai sensitivitas yang lebih baik dibandingkan tanpa menerapkan metode penyeimbangan data. Hal tersebut berdasarkan pada nilai rataan dan median sensitivitas yang bernilai diatas 0.87, meskipun nilai rataan dan median spesifisitas dan akurasi menjadi lebih kecil. Metode KNN dengan k = 6 disertai penanganan ketakseimbangan dengan metode ROS dipilih sebagai skema klasifikasi dengan KNN terbaik dibandingkan skema lainnya dalam memprediksi potensi keberhasilan studi mahasiswa program magister IPB dengan nilai rataan dan median sensitivitas sebesar 0.89 dan 0.89, nilai rataan dan median spesifisitas sebesar 0.75 dan 0.75, serta nilai rataan dan median akurasi sebesar 0.77 dan 0.77.

B. Saran

Penelitian ini masih memiliki beberapa kekurangan, baik dalam hal peubah yang digunakan, peubah yang paling berpengaruh terhadap respon, serta proses komputasi yang lama. Saran untuk penelitian selanjutnya agar menggunakan peubah peubah yang diharapkan mampu meningkatkan hasil klasifikasi KNN, seperti asal daerah perguruan tinggi S1, akreditasi jurusan S1, dan akreditasi perguruan tinggi S1, serta menerapkan metode yang dapat meningkatkan kinerja analisis klasifikasi KNN, seperti metode algoritma genetik untuk penentuan k optimum dan *K D tree* untuk mempersingkat proses perhitungan jarak data latih dan data uji. Selain itu, peubah respon sebaiknya terbagi kedalam kelas yang lebih rinci dalam menjelaskan status keberhasilan atau kegagalan studi mahasiswa, seperti lulus, mengundurkan diri, *dan drop out* (DO).

Daftar Pustaka

Garcia, V., J. Sanchez, and R. Mollineda. Exploring the performance of resampling strategies for the class imbalance problem. In *23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27(4)*, 857871.

Gu, Q., X. Wang, Z. Wu, B. Ning, and C. Xin (2016). An improved smote algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management 14(2)*, 92103.

Houben, I., L. Wehenkel, and M. Pavella (1997). Genetic algorithm based k nearest neighbors. *Journal IFAC Control of Industrial Systems 30(6)*, 1075–1080.

Indrayanti, D. Sugianti, and M. Karomi (2017). Peningkatan akurasi algoritma knn dengan seleksi fitur gain ratio untuk klasifikasi penyakit diabetes mellitus. *Jurnal IC-Tech 12(2)*, 1–6.

Irawan, E. (2015). Penggunaan random under sampling untuk penanganan ketidakseimbangan kelas pada prediksi cacat software berbasis neural network. *Journal of Software Engineering 1(2)*, 92–100.

Karno. Penentuan parameter pada algoritma klasifikasi k-nearest neighbor berbasis algoritma genetika. In *Seminar Nasional Teknologi Informasi (SNTI) 8 UNTAR*.

Kuramochi, M. and G. Karypis (2005). Gene classification using expression profiles: A feasibility study. *Int J Artif Intell Tools 14(4)*, 641660.

Ngai, E., Y. Hu, Y. Wong, Y. Chen, and X. Sun (2011). The application of data mining techniques in financial fraud detection: A classic cation framework and an academic review of literature. *Decision Support Systems 50(3)*, 559569.

Parvin, H., H. Alizadeh, and B. Bidgoli. Mknn: Modified k-nearest neighbor. In *Proceedings of the Word Congress on Engineering and Computer Science 2008 (WCECS 2008)*.

Saifudin, A. and R. Wahono (2015). Penerapan teknik ensemble untuk menangani ketidakseimbangan kelas pada prediksi cacat software. *Journal of Software Engineering 1(1)*, 28–37.

Siringoringo, R. (2017). Integrasi metode resampling dan k-nearest neighbor pada prediksi cacat software aplikasi android. *Jurnal ISD 2(1)*, 47–58.

Siringoringo, R. (2018). Klasifikasi data tidak seimbang menggunakan algoritma smote dan k-nearest neighbor. *Jurnal ISD 3(1)*, 44–49.

Wu, X., V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Stein-

berg (2008). Top 10 algorithms in data mining. *Journal of Knowledge and Information Systems 14(1)*, 1–37.

Yap, B., K. Rani, H. Rahman, S. Fong, Z. Khairudin, and N. Abdullah. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*.

Yu, D., J. Hu, Z. Tang, H. Shen, J. Yang, and J. Yang (2013). Neurocomputing improving protein-atp binding residues prediction by boosting svms with random under-sampling. *Neurocomputing 104(1)*, 180190.