METODE KLASIFIKASI DATA MINING DAN TEKNIK SAMPLING *SMOTE* MENANGANI *CLASS IMBALANCE* UNTUK SEGMENTASI CUSTOMER PADA INDUSTRI PERBANKAN

Hairani[*], Noor Akhmad Setiawan, Teguh Bharata Adji
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Jalan Grafika No. 2, Yogyakarta, 55281 Indonesia
[*]Email: ronii.ahmad83@gmail.com

Abstrak

Class imbalance merupakan sebuah permasalahan yang lazim ditemukan pada dataset, dimana disribusi antara class mayoritas (Negative) dan minoritas (positive) tidak seimbang. Dengan kata lain, class mayoritas memiliki jumlah yang lebih banyak dibandingkan class minoritas. Dengan distribusi yang tidak seimbang, metode pada machine learning cenderung keliru mengklasifikasikan class minoritas. Paper ini mengadopsi pendekatan teknik sampling yaitu Algoritma SMOTE untuk menangani permasalahan class imbalance yang dikombinasikan dengan metode klasifikasi yang lainnya yaitu metode J48, SVM, dan Naive Bayes. Berdasarkan hasil pengujian yang telah dilakukan dengan tools weka menggunakan evaluasi kinerja confusion matrix, menunjukkan bahwa metode J48+SMOTE memiliki tingkat akurasi dan sensitivity paling tinggi yaitu sebesar 0,93% dan 0,93%. Sedangkan metode SVM memiliki nilai specificity yang paling tinggi sebesar 0.99% dan metode Naive Bayes memiliki waktu komputasi yang paling cepat dibandingkan ketiga metode lainnya sebesar 0.38 seconds. Dengan demikian, metode J48+SMOTE mampu menangani class imbalance pada dataset Bank Direct Marketing pada industri perbankan dibandingkan metode SVM dan Naive Bayes.

***Kata kunci**: Algoritma SMOTE; Class Imbalance; Metode Klasifikasi*

I. PENDAHULUAN

Suatu *class* pada *dataset* dengan pendistribusian *class* yang tidak seimbang (*class imbalance*) menimbulkan kejadian klasifikasi lebih condong ke *class* mayoritas dibandingkan dengan *class* minoritas. Ketidakseimbangan *class* pada sebuah *dataset* merupakan suatu permasalahan dalam *machine learning*, dimana jumlah class *mayoritas (negative)* lebih besar dari pada jumlah *class minoritas (positive)*. Dengan kata lain, jumlah *class negative (mayoritas)* lebih besar jumlahnya dibandingkan dengan *class positive (minoritas)*. Permasalahan *class imbalance* merupakan sebuah permasalahan yang sudah lazim ditemukan pada dataset di berbagai bidang, termasuk prediksi cacat software, deteksi tumpahan minyak dari citra satelit, deteksi penipuan kartu kredit online, dan diagnosis penyakit(Phoungphol, 2013).

Masalah class imbalance biasanya terjadi ketika melakukan klasifikasi, dimana sebuah classifier cenderung mengklasifikasikan class mayoritas dan mengabaikan class minoritas. Untuk mengatasi permasalahan tersebut dapat digunakan dua pendekatan yaitu, pendekatan sample dan algoritma (Chawla et al. 2002).

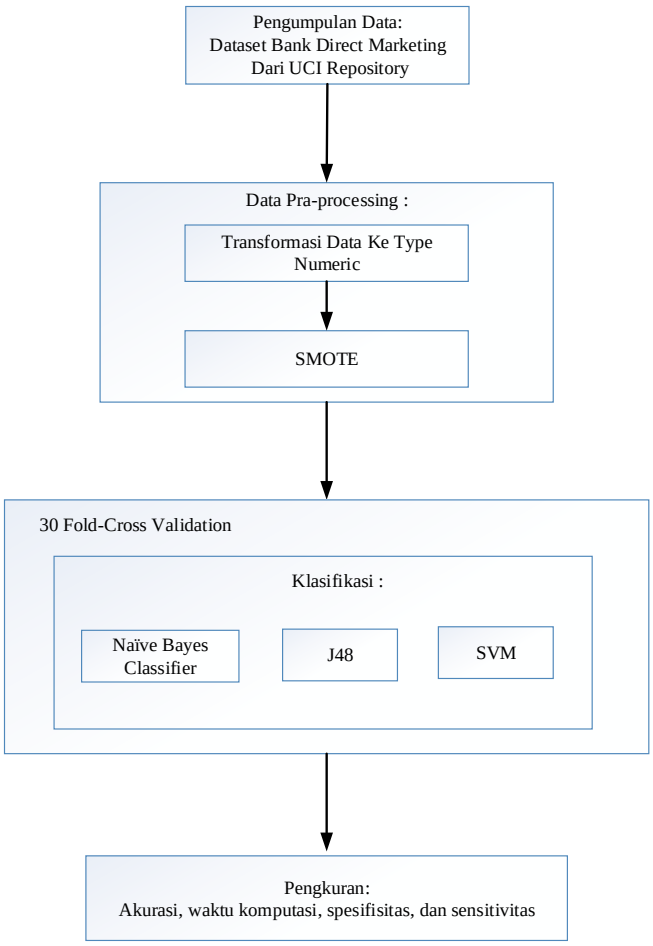
Sebagai contoh sebuah dataset yang tidak seimbang memiliki rasio 1:100, dimana 1 merepresentasikan *class minoritas (positive)* sedangkan 100 merepresentasikan *class mayoritas (negative)*. Sebuah metode klasifikasi yang mencoba untuk memaksimalkan akurasinya, dapat mencapai akurasi 99% hanya menggunakan *class negative (mayoritas)* tanpa melihat *class positive (minoritas)*. Hal tersebut dapat mengakibatkan metode pada *machine learning* cenderung keliru mengklasifikasikan yang seharusnya class minoritas dianggap sebagai class mayoritas.

Untuk mengatasi permasalahan *class imbalance*, dapat digunakan sebuah pendekaatan teknik sampling. Teknik sampling yang umumnya digunakan dalam mengatasi permasalahan *class imbalance* yaitu *Over-sampling*, *Under-sampling*, dan kombinasi keduanya. Penyelesaian class imbalance dilihat berdasarkan akurasi, *sensitivity*, dan *spesificity*dengan metode klasifikasi (J48, SVM, dan Naive Bayes) yang dikombinasikan algoritma SMOTE pada *dataset Bank Direct Marketing.*

ISBN 978-602-99334-5-1

II. METODOLOGI

Metodologi yang digunakan dalam penelitian ini dijelaskan pada Gambar 1 sebagai berikut:



Gambar 1. Metodologi Penelitian

Mengacu pada Gambar 1 tahapan awal penelitian dimulai dengan pengumpulan *dataset Bank Direct Marketing* yang diambil dari *UCI Repository*. *Dataset Bank Direct Marketing* memiliki 17 atribut yang ditunjukkan pada Tabel 1.

Tabel 1. Atribut Dataset Bank Direct Marketing

No	Atribut	Type	Value
1.	Age	Numeric	Real
2.	Job	Categorical	Admin, Unknown, Unemployed, Mangement, Housemaid, Entrepreneur, Stident, Blue-collar, Self-employed, Retired, Technician, Services
3.	Marital	Categorical	Married, Divorced, Single
4.	Education	Categorical	Unknown, Secondary, Primary, Tertiary
5.	Default	Binary	Yes or No

6.	Balance	Numeric	Real
7.	Housing	Binary	Yes or No
8.	Loan	Binary	Yes or No
9.	Contact	Categorical	Unknown, Telephone, Cellular
10.	Day	Numeric	Real
11.	Month	Categorical	Jan, Feb, Mar,……, Nov, Dec
12.	Duration	Numeric	Real
13.	Campaign	Numeric	Real
14.	Pdays	Numeric	Real
15.	Previous	Numeric	Real
16.	Poutcome	Numeric	Real
17.	Class	Categorical	Yes or No

Pada tahapan *pre-processing*, type data pada *dataset bank direct marketing* ditransformasikan ke bentuk numerik. Kemudian dilakukan klasifikasimenggunakan metode klasifikasi Naive Bayes, J48, dan SVM yang divalidasi dengan 30 cross validation untuk mendapatkan model yang maksimal, yang diukur berdasarkan tingkat akurasi, *sensitivity*, *specificity*, dan waktu komputasi yang dibutuhkan dalam membangun sebuah model klasifikasi.

III. HASIL DAN PEMBAHASAN

Hasil pengujian yang telah dilakukan menggunakan 30 Fold-cross validation dan evaluasi kinerjanya menggunakan *confusion matrik* yang diukur berdasarkan akurasi, *sensitivity*, *specificity*, dan waktu komputasi yang masing-masing rumusnya ditunjukkan pada persaman 1, 2, dan 3. Adapun confusion matrix nya ditunjukkan pada Tabel 2.

Tabel 2. Confusion Matrix

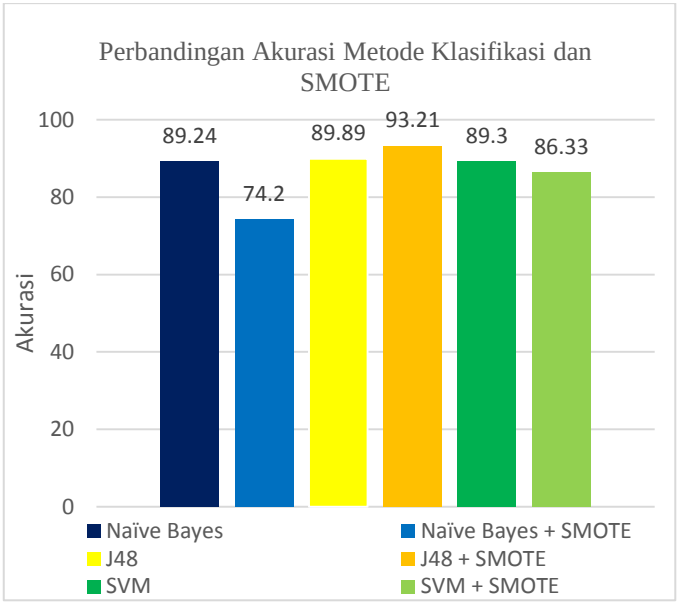
		Class Prediksi	
		Yes	**No**
Class Actual	**Yes**	TP	FN
	No	FP	TN

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

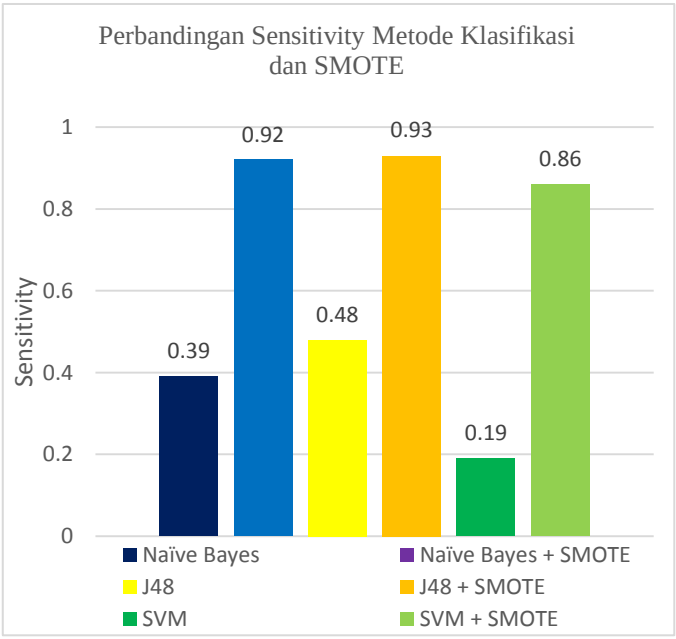
$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

ISBN 978-602-99334-5-1



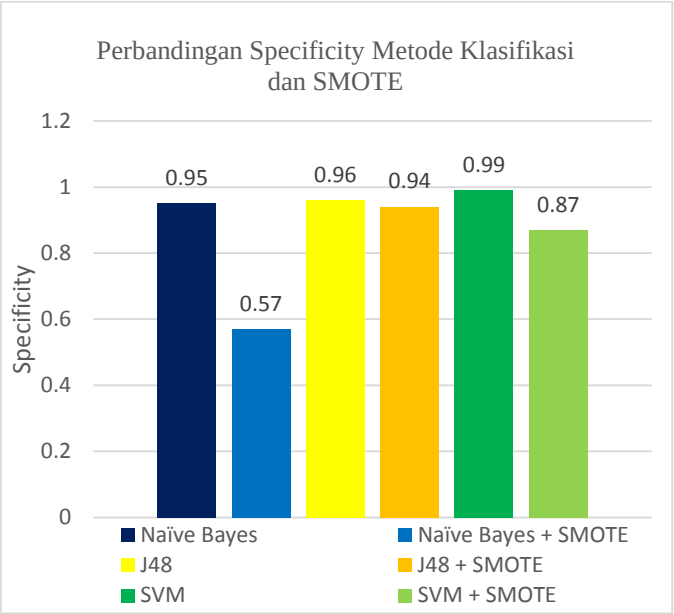
Gambar 2. Hasil Perbandingan Akurasi Metode Klasifikasi dan SMOTE

Berdasarkan hasil klasifikasi yang ada pada Gambar 2 menunjukkan bahwa Metode J48+SMOTE memiliki akurasi yang paling tinggi dibandingkan dengan metode laiinya yaitu sebesar 93,21%, disusul oleh metode J48, SVM, Naive Bayes, SVM+SMOTE, dan Naive Bayes SMOTE masing-masing akurasinya sebesar 89,89%, 89,3%, 89,24%, 86,33%, dan 74,2%.



Gambar 3 Hasil Perbandingan Sensitivity Metode Klasifikasi dan SMOTE

Berdasarkan nilai *sensistivity* yang ada pada Gambar 3 menunjukkan bahwa Metode J48+SMOTE memiliki nilai *sensistivity* yang paling tinggi dibandingkan dengan metode laiinya sebesar 93%, disusul oleh metode Naive Bayes+SMOTE, SVM+SMOTE, J48, Naive Bayes, dan SVM masing-masing akurasinya sebesar 92%, 86%, 48%, 39%, dan 19%.



Gambar 4 Hasil Perbandingan Specificity Metode Klasifikasi dan SMOTE

Berdasarkan nilai *spesificity* yang ada pada Gambar 4 menunjukkan bahwa Metode SVM memiliki nilai *spesificity* yang paling tinggi dibandingkan dengan metode lainnya sebesar 99%, disusul oleh metode J48, Naive Bayes, J48+SMOTE, SVM+SMOTE dan Naive Bayes+SMOTE masing-masing akurasinya sebesar 96%, 95%, 94%, 87%, dan 57%

Tabel 3 Waktu Komputasi (*Seconds*)

Metode	Data Original	SMOTE
Naive Bayes	0,38	0,66
J48	14,92	19,14
SVM	1662,55	1592,52

Berdasarkan waktu komputasi yang dibutuhkan pada Tabel 3 menunjukkan bahwa Metode Naïve Bayes memiliki waktu komputasi yang paling cepat dibandingkan dengan metode laiinya sebesar 0.38 *seconds*, disusul oleh metode Naive Bayes+SMOTE, J48, J48+SMOTE, SVM+SMOTE dan SVM masing-masing waktu komputasinya sebesar 0.66 *seconds*, 14,92 *seconds*, 19,14 *seconds*, 1592,52 *seconds*, dan 1662,55 *seconds*.

IV. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan dengan menggunakan *tools weka* yang di evaluasi kinerjanya menggunakan confusion matrix. Didapatkan hasil penelitian bahwa metode J48+SMOTE memiliki tingkat akurasi dan *sensisitivity* yang paling tinggi dibandingkan dengan metode lainnya, dari segi *specificity* metode SVM memiliki nilai paling tinggi, dan segi komputasi waktu metode Naive Bayes memiliki waktu yang paling cepat dibandingkan dengan metode lainnya.Dengan demikian, metode J48+SMOTE mampu menangani class imbalance pada dataset Bank Direct Marketing pada industri perbankandibandingkan metode SVM dan Naive Bayes.

DAFTAR PUSTAKA

Chawla, N. V, Bowyer, K.W. & Hall, L.O., 2002. SMOTE : Synthetic Minority Over-sampling TEchnique. , 16.

Phoungphol, P., 2013. *A Classification Framework for Imbalanced Data*. Georgia State University. Available at: http://scholarworks.gsu.edu/cs_diss/78.

ISBN 978-602-99334-5-1