JURNAL INFORMATIKA, Vol. 10 No. 2 Oktober 2023, Halaman 101-108
ISSN: 2355-6579 | E-ISSN**:** 2528-2247
DOI: https://doi.org/10.31294/inf.v10i2.15540

Perbandingan Metode Klasifikasi pada Data dengan *Imbalance Class* dan *Missing Value*

Nofita Istiana[1*], Arief Mustafiril[2]

[1] Politeknik Statistika STIS
Jalan Otto Iskandardinata No. 64C, Jakarta 13330, Indonesia

[2] Universitas Trisakti
Jalan Kyai Tapa No. 1 Grogol, Jakarta Barat, Indonesia

e-mail: [1]nofita@stis.ac.id, [2]firil.leyan@gmail.com

Informasi Artikel	Diterima: 02-03-2023	Direvisi: 03-06-2023	Disetujui: 11-07-2023

Abstrak

Imbalance class dan *missing value* merupakan beberapa permasalahan dalam metode klasifikasi. *Imbalance class* mengakibatkan prediksi kelas minoritas disalahklasifikasikan menjadi kelas mayoritas. *Missing value* menyebabkan beberapa algoritma dalam metode klasifikasi tidak dapat dijalankan. Pada penelitian ini, *imbalance class* ditangani dengan SMOTE, sedangkan *missing value* ditangani dengan imputasi rataan dan *binning* peubah. Metode klasifikasi yang dibandingkan dalam penelitian ini adalah *logistic regression*, *bagging*, *boosting*, *random forest*, dan *support vector machine* yang diaplikasikan pada data *dummy* status kolektibilitas debitur berjumlah 12459. Data tersebut berisi debitur berstatus kolektibilitas baik sebanyak 97.48 dan debitur berstatus kolektibilitas buruk sebanyak 2.52 persen. Metode yang memberikan akurasi tertinggi yaitu *random forest* (*missing value* diimputasi dengan nilai rataan), yang menghasilkan akurasi sebesar 80.1 persen, sensitivitas sebesar 59.3 persen, dan spesifisitas sebesar 80.7 persen.

Kata Kunci: metode klasifikasi, SMOTE, *Weight of Evidence* (WoE)

Abstract

Imbalance class and missing value are some of the problems in classification method. Imbalance class causes the prediction of the minority class to be misclassified as the majority class. Missing value causes several algorithms in classification method cannot be run. In this study, imbalance class is handled by SMOTE, while missing value is handled by mean imputation and binning variable. The classification methods being compared in this study are logistic regression, bagging, boosting, random forest, and support vector machines which are applied to dummy data on debtors' collectibility status with total data 12459. The data contains 97.48 debtors with good collectibility status and 2.52 percent of debtors with bad collectibility status. The method that provides the highest accuracy is random forest (missing value imputed by mean value), which results in accuracy of 80.1 percent, sensitivity of 59.3 percent, and specificity of 80.7 percent.

Keywords*: classification method, SMOTE*, *Weight of Evidence* (WoE)

1. Pendahuluan

Metode klasifikasi merupakan suatu teknik dalam *data mining* untuk menentukan objek ke dalam kelompok/kategori yang telah ditentukan menurut karakteristik dari objek tersebut (Wibawa & Dkk, 2018). Metode ini banyak digunakan dalam proses pengambilan keputusan yang cerdas (*intelligent decision making*). Metode klasifikasi digunakan dalam berbagai bidang, salah satu contohnya yaitu dalam bidang perbankan, misalnya ketika memprediksi status kolektibilitas dari debitur. Debitur merupakan perusahaan yang memperoleh pinjaman/kredit. Kolektibilitas/ kualitas kredit merupakan kemampuan debitur untuk mengembalikan dana yang dipinjam dari bank baik pinjaman pokok maupun bunga kreditnya pada waktu yang telah ditentukan berdasarkan perjanjian yang telah disepakati (Dinaloni & Putri, 2018).

Salah satu permasalahan dalam metode klasifikasi adalah ketidakseimbangan kelas (*imbalance class*). *Imbalance class* adalah kondisi di mana suatu himpunan data terdapat

satu kelas yang memiliki jumlah yang jauh lebih kecil dibanding kelas lainnya. *Imbalance class* menyebabkan kelas minoritas dipredikasi secara salah menjadi kelas mayoritas (N & Sudaryanto, 2022). Selain itu, *imbalance class* mengakibatkan tingkat sensitivitas rendah pada data dengan dua kategori atau binomial. Data kolektibilitas debitur dalam penelitian ini kemungkinan besar terjadi *imbalance class* yaitu debitur dengan status kolektibilitas baik (*good*) lebih besar dibandingkan dengan status kolektibilitas buruk (*bad*).

Masalah lain yang dapat muncul yaitu banyaknya data kosong (*missing value*). *Missing value* dapat menyebabkan beberapa algoritma dalam metode klasifikasi tidak dapat dijalankan. *Missing value* dapat terjadi karena karakteristik tentang observasi tidak ditemukan, sulit didapatkan atau memang tidak ada informasi mengenai observasi tersebut. Hal ini mengakibatkan tingkat akurasi rendah dan kualitas data yang buruk saat data diolah. Salah satu metode untuk menanggulangi *missing value* adalah dengan mengganti nilai yang tidak ada dengan suatu nilai yang diasumsikan sesuai, atau disebut sebagai *imputation*/ imputasi (Mawarsari, 2016).

Data dengan *imbalance class* dan banyaknya *missing value* perlu diberikan perlakuan tertentu agar diperoleh metode klasifikasi dengan ukuran ketepatan prediksi yang tinggi. *Imbalance class* dapat diatasi dengan memanipulasi data, memodifikasi algoritma, atau gabungan keduanya (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Cara paling mudah yaitu *resampling* data (Kotsiantis et al., 2006). Beberapa teknik *resampling* yang sering digunakan yaitu *oversampling*, *undersampling*, dan SMOTE (*Synthetic Minority Over-samplingTechnique).* Penerapan SMOTE dapat meningkatkan sensitivitas hingga lebih dari 50 persen meskipun akurasi dan spesifitasnya menurun dibandingkan tanpa SMOTE (Wijaya, Soleh, & Rizki, 2018). Sedangkan menurut Nikmatul Kasanah, Muladi, & Pujianto (2017), penggunaan SMOTE dapat menghasilkan akurasi yang lebih tinggi dibandingkan tanpa SMOTE. *Missing value* dapat ditangani dengan imputasi data atau melakukan *binning* (pengkategorian) pada peubah prediktor, sehingga *missing value* menjadi satu kategori tersendiri. Penelitian mengenai metode klasifikasi saat ini hanya fokus menangani *imbalance class* saja, seperti penelitian Nikmatul Kasanah et al. (2017), Wijaya et al. (2018), dan Astuti & Lenti (2021). Penelitian yang bertujuan untuk menangani secara sekaligus masalah *imbalance class* dan *missing value* masih terbatas. Oleh karena itu, penulis melakukan perbandingan metode dalam penelitian ini.

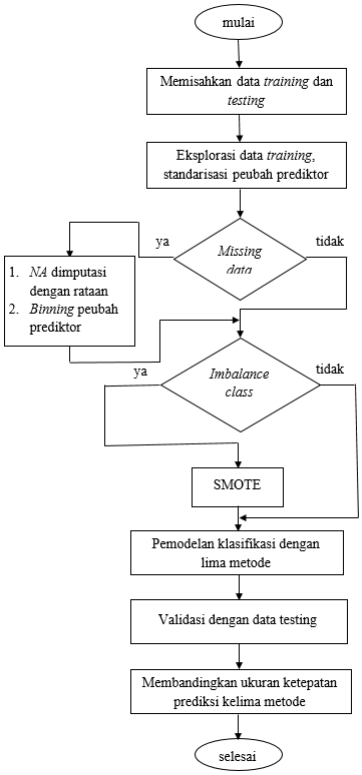
Metode klasifikasi yang dibandingkan dalam penelitian ini adalah *logistic regression*, *bagging*, *boosting* dengan GBM *(Generalized Boosting Models)*, *random forest*, dan *support vector machine*. Metode tersebut dipilih karena berdasarkan penelitian sebelumnya yang dilakukan oleh Wibowo (2015), Aulia, Hadiyoso, & Ramadan (2015), Prasetio & Pratiwi (2015), serta N & Sudaryanto (2022), metode tersebut yang menghasilkan akurasi cukup tinggi. Selain itu, akan dibandingkan pula hasil metode klasifikasi tersebut dengan penggunaan imputasi data dan *binning* dalam penanganan *missing value*.

2. Metode Penelitian

Penelitian ini menggunakan data *dummy* yang merupakan rekaan 12459 debitur. Data tersebut berisi status kolektibilitas dan peubah-peubah lain merupakan rasio yang menggambarkan kualitas perusahaan, seperti rasio profit terhadap aset, rasio hutang terhadap aset, dan lain-lain. Untuk alasan *confidentiality*, tidak disebutkan nama peubahnya dalam data. Data dapat diakses pada tautan https://s.stis.ac.id/dummydebitur atau https://bit.ly/datadummydebitur.

Terdapat lima metode yang akan dibandingkan tingkat akurasinya pada penelitian ini, yaitu *logistic regression*, *bagging*, *boosting* dengan GBM *(Generalized Boosting Models)*, *random forest*, dan SVM*.* Penelitian ini menggunakan *software* R dalam semua tahapannya.

Flowchart langkah kerja (Gambar 1) menampilkan tahapan yang dilakukan dalam penelitian ini. Tahapan awal yang dilakukan yaitu memisahkan data *training* dan data *testing*. Selanjutnya melakukan eksplorasi data *training* untuk mengetahui gambaran awal tentang data yang digunakan. Peubah prediktor memiliki rentang yang sangat bervariasi, sehingga dilakukan standarisasi data. Kemudian melakukan penanganan *missing value* dengan dua cara yaitu *missing value* dimputasi dengan rataan dan *binning* peubah prediktor. Selanjutnya dilakukan penanganan *imbalance class* dengan SMOTE. Metode klasifikasi dilakukan dengan lima metode seperti yang dijelaskan sebelumnya. Kemudian dilakukan validasi dengan data *testing* dan membandingkan ukuran ketepatan prediksi dengan cara menghitung rataan geometrik terhadap akurasi, sensitivitas, dan spesifisitas yang dihasilkan dari kelima metode.



Sumber: Penulis (2023)

Gambar 1. *Flowchart* Langkah Kerja

2.1 *Imbalance Class*

Imbalance class (klasifikasi kelas tidak seimbang) adalah suatu masalah ketimpangan dalam distribusi data yang signifikan, yaitu salah satu kelas mempunyai sangat banyak jumlah data (mayoritas) dan kelas yang lain mempunyai sangat sedikit jumlah data (minoritas) (Siringoringo, 2018). Metode klasifikasi biasa membuat kelas minoritas tidak dapat diprediksi dengan baik. Hal ini dikarenakan data pada satu kelas memiliki jumlah yang sangat kecil sehingga kelas tersebut akan diprediksi menjadi kelas mayoritas. Salah satu cara untuk menanggulangi *imbalance class* yaitu dengan SMOTE.

2.2 *Synthetic Minority Over-sampling Technique* (SMOTE)

SMOTE adalah metode yang dapat digunakan dalam menanggulangi *imbalance class*. Metode ini membuat *instance* baru yang berasal dari *minority class* agar data menjadi seimbang. SMOTE memiliki ide pokok yaitu menambahkan jumlah sampel dalam *minority class* agar setara dengan *majority class* dengan cara membangkitkan data baru (*synthetic data*) menurut tetangga terdekat/*k-nearest neighbor* (Wijaya et al., 2018).

2.3 *Missing Value*

Missing value (data hilang) merupakan sebagian atau seluruh informasi yang hilang di suatu observasi dalam data. Metode yang dapat digunakan untuk menangani *missing value* adalah dengan imputasi (*imputation*). Imputasi dilakukan dengan cara mengganti nilai *missing value* dengan nilai tertentu berdasarkan karakteristik data (Hendrawati, 2015). Akan tetapi, jika terlalu banyak nilai yang diimputasi maka dikhawatirkan data tidak lagi mencerminkan data yang sesungguhnya. Menurut Juhola & Laurikkala (2013), untuk dataset dua kelas, meskipun proporsi data hilang pada suatu kasus sebesar 20–30 persen, hasilnya hampir sama baiknya dengan tanpa data hilang hilang. Jika ada lebih dari dua kelas maka proporsi data hilang sebesar 10-20 persen mungkin terlalu banyak, setidaknya untuk kelas kecil dengan jumlah yang relatif sedikit.

Cara lain untuk menangani *missing value* yaitu dengan *binning* peubah, sehingga *missing value* dikelompokkan menjadi kategori tersendiri. Salah satu teknik *binning* yaitu *Weight of Evidence* (WoE). WoE merupakan kemampuan prediksi dari peubah prediktor dalam hubungannya dengan peubah respon. Manfaat WoE yaitu membantu mengubah peubah prediktor yang berbentuk kontinyu menjadi beberapa grup atau bin berdasarkan kesamaan peubah respon. Rumus WoE yaitu sebagai berikut.

$$WoE = ln\left(\frac{\%\ non-events}{\%\ events}\right) \qquad (1)$$

% non-events adalah status kolektibilitas baik dan % events adalah status kolektibilitas buruk yang ada di tiap grup. WoE menghasilkan suatu nilai yang disebut *Information Value* (IV). Sifat peubah prediktor yang digunakan ditentukan dari nilai IV (Tabel 1). IV dapat digunakan untuk mengurutkan peubah berdasarkan kepentingannya, sehingga IV dapat digunakan untuk seleksi peubah. Rumus IV yaitu sebagai berikut.

$$IV = \sum(\%non\ events - \%\ events)xWoE \qquad (2)$$

Tabel 1. Kriteria *Information Value* Peubah Prediktor

Information Value	Sifat peubah prediktor
Kurang dari 0.02	Tidak cocok untuk prediksi
0.02 s.d. 0.1	Kekuatan prediksi lemah
0.1 s.d. 0.3	Kekuatan prediksi sedang
0.3 s.d. 0.5	Kekuatan prediksi kuat
>0.5	Kekuatan prediksi mencurigakan

2.4 *Logistic Regression*

Logistic regression diterapkan dalam melihat hubungan peubah respon berbentuk *dichotomous data*/ data biner dengan peubah prediktornya. *Logistic regression* mirip dengan *linear regression*, tetapi peubah responnya bernilai 0 dan 1. Dalam penelitian ini, peubah respon adalah 0 jika kolektibilitas debitur berkategori baik dan 1 jika kolektibilitas debitur berkategori buruk. Model *logistic regression* menghasilkan sutu nilai peluang sukses dari observasi. Perasamaan model *logistic regression* adalah sebagai berikut (Agresti, 2002):

$$\pi(x) = \frac{e^{(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}} \qquad (3)$$

$\pi(x)$ merupakan peluang sukses dengan nilai 0≤ $\pi(x)$ ≤1, β merupakan nilai parameter dengan *p* adalah banyaknya peubah prediktor. Transformasi logit terhadap $\pi(x)$ yang non linier diperlukan untuk memperoleh fungsi yang linier, sehingga hubungan peubah prediktor dan peubah respon dapat terlihat.

2.5 *Bagging*

Bagging merupakan salah satu metode dalam *machine learning* yang bertujuan memperbaiki hasil dari algoritma klasifikasi (Prasetio & Pratiwi, 2015). *Bagging* dapat dilakukan untuk mengurangi variansi dari suatu prediktor dalam metode klasifikasi dan regresi. *Bagging* bermanfaat untuk memperbaiki kualitas prediksi. *Bagging* juga dapat menanggulangi ketidakstabilan dari *single classification method*. Jumlah replikasi *bootstrap* yang digunakan memengaruhi tingkat ketepatan klasifikasi dari *bagging*. Oleh karena itu, kebaikan *bagging* dipengaruhi oleh penentuan banyaknya replikasi *bootstrap*.

Bagging merupakan singkatan dari Bootstrap Aggregating. *Bootstrap* adalah penarikan data sampel yang independen (*resampling*) dan berulangkali dilakukan untuk memperkirakan tingkat kesalahan (*error*) dari pengulangan tersebut. Sampel *bootstrap* diambil dari sampel asli secara random dengan pemulihan (*with replacement*). Metode *ensemble* ini memiliki ide dasar yaitu *resampling* secara random dengan pengembalian pada data latih (*training*). Gugus data/data *training* yang baru dihasilkan untuk membangkitkan pengklasifikasi dengan banyak bentuk. *Aggregating* merupakan penggabungan beberapa nilai prediksi menjadi satu nilai prediksi (Zhou, 2012). *Aggregating* pada *bagging* memiliki konsep yaitu pemilihan (*voting*) dalam kasus klasifikasi dan rataan dalam kasus regresi.

Proses prediksi secara *bagging* yaitu sebagai berikut:

a. Tahap *bootstrap*
   - Mengambil sampel random *with replacement* dari data *training*.
   - Menyusun pohon terbaik data tersebut.
   - Mengulangi langkah kesatu dan kedua sehingga diperoleh buah pohon klasifikasi sebanyak *n*.

b. Tahap *aggregating*
   - Melakukan prediksi gabungan berdasarkan *n* buah pohon klasifikasi tersebut dengan aturan suara terbanyak (*majority vote*).

2.6 *Boosting*

Boosting serupa dengan *bagging* yaitu beberapa model digunakan dengan algoritma yang sama. Setiap model dalam *bagging* memiliki bobot (*weight*) yang sama. Sedangkan setiap model dalam *boosting* smemiliki bobot yang berbeda (Arrahimi, Ihsan, Kartini, Faisal, & Indriani, 2019). *Voting* dengan bantuan bobot dari masing-masing model menentukan prediksi akhir. Hal inilah yang mengakibatkan teknik *boosting* mampu memperbaiki kesalahan prediksi dari model sebelumnya.

2.7 *Random Forest*

Random forest (RF) dilakukan untuk memperbaiki proses prediksi dengan *bagging*. Adanya tahap *random sub-setting* ketika pembentukan *tree* merupakan perbedaan pokok dari kedua metode ini (Wibowo, 2015). Tahapan RF adalah sebagai berikut:

a. Tahap *bootstrap* yaitu menarik sampel random *with replacement* berukuran *n* dari data *training*.
b. Tahap *random sub-setting* yaitu menyusun pohon (*tree*) berdasarkan sampel tersebut, tetapi di setiap proses pemisahan dipilih random *m* < *d* peubah prediktor, dan melakukan pemisahan yang paling baik.
c. Melakukan *bootstrap* dan r*andom subsetting* sebanyak *k* kali sehingga diperoleh *k* pohon random.
d. Melakukan prediksi gabungan dengan *majority vote* dalam kasus klasifikasi atau rataan dalam kasus regresi berdasarkan *k* buah pohon tersebut.

Proses menggabungkan nilai prediksi dari beberapa pohon yang dihasilkan mirip dengan *bagging*. Seluruh peubah prediktor yang ada tidak dipakai seluruhnya untuk melakukan pemisahan tapi hanya sebagian hasil pemilihan secara random di setiap pembentukan pohon. Sehingga kumpulan pohon tunggal dihasilkan dengan ukuran dan bentuk yang berbeda. Harapannya yaitu kumpulan pohon tunggal memiliki korelasi yang kecil antar pohonnya. Korelasi kecil menyebabkan variasi prediksi dari

RF menjadi kecil dan lebih kecil dibanding variasi prediksi *bagging.*

2.8 *Support Vector Machine (SVM)*

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik pada tahun 1992. Metode ini merupakan salah satu metode favorit dalam *pattern recognition*. Ide pokok SVM (Aulia et al., 2015) adalah pengklasifikasi linier di ruang kerja berdimensi tinggi yang dikembangkan untuk mengatasi *problem non-linear* dengan cara memasukkan konsep *kernel trick*. Pasangan data masukan dan data keluaran berupa sasaran yang diinginkan digunakan dalam pembelajaran. Pembelajaran dengan cara ini disebut dengan pembelajaran terarah (*supervised learning*). *Supervised learning* ini menghasilkan fungsi yang menggambarkan bentuk ketergantungan input dan outputnya. Fungsi yang diperoleh diharapkan mempunyai kemampuan generalisasi yang baik. Artinya fungsi tersebut dapat digunakan untuk data masukan di luar data pembelajaran.

2.9 Penilaian Kebaikan Klasifikasi

Untuk membandingkan kinerja dari *logistic regression*, *bagging*, *boosting* dengan GBM *(Generalized Boosting Models)*, *random forest*, dan SVM dapat dilihat dari tingkat akurasi, sensitivitas, dan spesifisitas. *Confusion matrix* digunakan untuk mengevaluasi kinerja dari berbagai metode tersebut. Informasi tentang hasil klasifikasi data aktual (keadaan sesungguhnya) dan data hasil prediksi termuat dalam *confusion matrix* (Tabel 2). Harapannya yaitu terdapat keselarasan antara data aktual dan data hasil prediksi. Jika data aktual status kolektibilitas termasuk kategori *good* maka diharapkan hasil prediksinya juga menghasilkan kategori *good*. Begitu pula sebaliknya.

Tabel 2. *Confusion Matrix*

		Data Hasil Prediksi	
		Good	*Bad*
Data aktual	*Good*	*True Positive*	*False Negative*
	Bad	*False Positive*	*True Negative*

Keadaan *True Positive* terjadi jika observasi yang sesungguhnya bersatus *good* juga diprediksi *good*. Keadaan *False Negative* terjadi jika observasi yang sesungguhnya berkategori *good* tetapi di prediksi *bad*. Keadaan *False Positive* terjadi jika observasi yang sesungguhnya berkategori *bad* tetapi di prediksi *good*. Keadaan *True Negative* terjadi jika observasi yang sesungguhnya berkategori *bad* juga diprediksi *bad*. Harapannya yaitu *False Negative* dan *False Positive* bernilai sekecil mungkin sehingga tingkat keakuratan (*accuracy*), kepekaan (*sensitivity*), dan kekhususan (*specificity*) menjadi setinggi mungkin. Rumus dalam penghitungan akurasi, sensitivitas, dan spesifisitas adalah sebagai berikut:

a. Akurasi

Akurasi adalah tingkat ketepatan prediksi secara menyeluruh, yaitu persentase banyaknya prediksi yang tepat di seluruh observasi.

$$\text{Akurasi} = \frac{n\,(true\ positive) + n\,(true\ negative)}{n\,(observasi)}\, x\ 100\% \quad (4)$$

b. Sensitivitas

Sensitivitas merupakan persentase ketepatan prediksi pada kelas positif, artinya observasi di kelas positif pada data aktual diprediksi positif (*true positive*).

$$\text{Sensitivitas} = \frac{n\,(true\ positive)}{n\,(true\ positive) + n\,(false\ negative)} \quad (5)$$

c. Spesifisitas

Spesifisitas adalah tingkat ketepatan pada kelas negatif dalam data aktual, yaitu persentase banyaknya prediksi yang tepat pada observasi yang sebenarnya negatif.

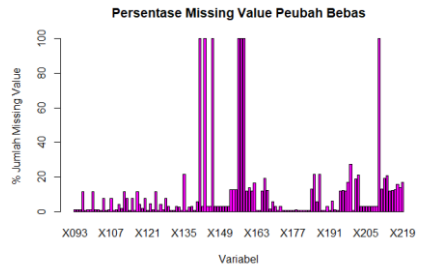
$$\text{Spesifisitas} = \frac{n\,(true\ negative)}{n\,(true\ negative) + n\,(false\ positive)} \quad (6)$$

n dalam rumus di atas menyatakan jumlah. Ketiga ukuran kebaikan klasifikasi tersebut diharapkan setinggi mungkin. Akurasi tinggi tanpa diikuti sensitivitas tinggi maka prediksinya menjadi kurang baik. Kinerja dari metode *logistic regression*, *bagging*, *boosting* dengan GBM *(Generalized Boosting Models)*, *randomForest*, dan SVM dievaluasi melalui nilai akurasi, sensitivitas, dan spesifisitas.

3. Hasil dan Pembahasan

3.1 Eksplorasi Data

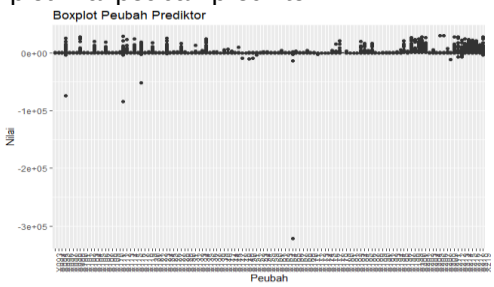
Data terdiri atas 12459 pengamatan dengan satu peubah respon (peubah Y) yaitu status kolektibilitas debitur dan 127 peubah prediktor (peubah X). Hasil eksplorasi data 12459 debitur menunjukkan bahwa terjadi *imbalance class* pada status kolektibilitas. Hal ini karena proporsi kategori *good* (0) dan *bad* (1) yaitu 97.48 persen dan 2.52 persen. Selain itu, data tersebut juga terdapat *missing value* yang cukup besar karena hampir seluruh peubah terdapat *missing value*.



Sumber: Penulis (2023)

Gambar 2. Persentase *Missing Value* Peubah Prediktor

Gambar 2 menunjukkan bahwa terdapat cukup banyak data hilang pada data. Bahkan, terdapat peubah yang seluruhnya tidak ada nilainya, sehingga dilakukan penghapusan terhadap peubah tersebut. Berikut ini adalah boxplot nilai peubah prediktor.



Gambar 3. Boxplot Peubah Prediktor Status Kolektibilitas Debitur

Dari boxplot di atas tampak cukup banyak peubah prediktor yang terdapat *outlier* (data pencilan). Sebagian besar peubah bernilai di sekitar nol, tetapi ada amatan yang nilainya sangat jauh dari pusat data.

3.2 Metode Klasifikasi

Untuk menangani *imbalance class*, dilakukan *resampling* data dengan SMOTE. Untuk menangani *missing value* dilakukan perbandingan imputasi dengan rataan dan *binning* peubah prediktor dengan *woebinning*. Pada imputasi tidak dilakukan seleksi peubah karena nama peubah tidak disebutkan sehingga peneliti tidak dapat melihat tingkat kepentingan (secara teori) dari masing-masing peubah prediktor terhadap peubah respon. Sedangkan, dalam *binning*, seleksi peubah dilakukan dengan melihat *information value* dari masing-masing peubah. *Information value* yang dipakai dalam seleksi peubah yaitu 0.1.

Pengolahan data dengan *logistic regression* dilakukan dengan menggunakan fungsi glm dengan *link function* binomial. Data yang digunakan diseimbangkan terlebih dahulu dengan SMOTE yang ada di library (DMwR). Hasil yang diperoleh dari model adalah nilai peluang debitur berstatus kolektibilitas buruk.

Pemodelan dengan *bagging* menggunakan fungsi bagging pada library (adabag). Prediksi yang dihasilkan model ini berbeda-beda tergantung dari *control* yang digunakan. *Control* yang digunakan harus mempertimbangkan *trade off* antara waktu pengolahan dan perubahan nilai akurasi yang dihasilkan. Metode *bagging* yang digunakan yaitu dengan control nbagg=50, minbucket=5, dan cp=0.005.

Pemodelan *boosting* dilakukan dengan menggunakan library gbm. Hasil akurasi pemodelan *boosting* sangat bergantung pada jumlah pohon. Pengolahan dengan jumlah pohon menghasilkan kesimpulan bahwa semakin besar jumlah pohon maka semakin tinggi akurasi dan spesifisitasnya, tetapi semakin rendah sensitivitasnya. Dalam penelitian ini, jumlah pohon yang digunakan yaitu seratus pohon. Sama halnya dengan *logistic regression*, output yang dihasilkan dari prediksi model *boosting* adalah nilai peluang debitur berstatus kolektibilitas buruk.

Pemodelan *random forest* dilakukan dengan menggunakan library (randomForest). Jumlah maksimal simpul pohon terminal (maxnodes) yaitu seratus dan jumlah pohon (ntree) sebanyak empat puluh.

Pada peneitian ini svm dijalankan dengan menggunakan library (e1071). Model yang dihasilkan melalui svm sangat dipengaruhi oleh tuning parameter yang dilakukan. Parameter tuning optimum dapat diperoleh dengan menjalankan fungsi tune.svm. Meskipun begitu pemilihan parameter yang akan di tuning juga sangat mempengaruhi. Jika dibandingkan dengan algoritma-algoritma sebelumnya, svm membutuhkan waktu komputasi yang jauh lebih lama. Berikut ini adalah rangkuman nilai akurasi dari kelima metode klasifikasi.

Tabel 3. Akurasi Kelima Metode Klasifikasi dengan Imputasi pada *Missing Value* (dalam 100 persen)

Metode Klasifikasi	Imputasi Terhadap *Missing Value*			Rataan
	Akurasi	Sensitivitas	Spesifisitas	Geometrik
Logistic regression	0.674078	0.674419	0.6740696	0.674189
Bagging	0.904037	0.302326	0.9195678	0.631076
Boosting (gbm)	0.855179	0.430233	0.8661465	0.683047
Random forest	0.801346	0.593023	0.8067227	0.726449
SVM	0.698069	0.616279	0.7001801	0.670340

Dari tabel 3 tampak bahwa jika data status kolektibilitas debitur yang *missing value* diimputasi dengan nilai rataan, akurasi tertinggi yaitu dengan metode *bagging*. Akan tetapi, metode *bagging* menghasilkan sensitivitas yang paling rendah dibanding kelima metode lainnya. Akurasi terendah yaitu metode *logistic regression*. Akan tetapi, *logistic regression* menghasilkan sensitivitas yang paling tinggi. Nilai spesifisitas tertinggi juga dihasilkan oleh metode *bagging*, sedangkan terendah dihasilkan oleh *logistic regression*. Metode klasifikasi yang dipilih yaitu yang seimbang antara akurasi, sensitivitas, dan spesifisitas. Sehingga untuk pemilihan metode yang akan digunakan untuk prediksi dilakukan penghitungan rataan geometrik terhadap akurasi, sensitivitas, dan spesifisitas. Dari tabel 3, rataan geometrik tertinggi yaitu dengan metode *random forest*. Sehingga *random forest* lebih baik dibanding metode lainnya jika dilakukan imputasi terhadap data yang hilang.

Tabel 4. Akurasi Kelima Metode Klasifikasi dengan WoE Pada Peubah Prediktor (dalam 100 persen)

No	Metode Klasifikasi	WoE *Binning*			Rataan Harmonik
		Akurasi	Sensitivitas	Spesifisitas	
1	*Logistic regression*	0.7790262	0.4117647	0.7875719	0.6321650
2	*Bagging*	0.8418941	0.2823529	0.8549138	0.5879278
3	*Boosting* (gbm)	0.8384163	0.3294118	0.8502601	0.6169504
4	*Random forest*	0.8349385	0.2705882	0.8480701	0.5764987
5	SVM	0.8188871	0.3294118	0.8302765	0.6072889

Dari tabel 4 tampak bahwa jika data status kolektibilitas debitur dilakukan *binning* (pengkategorian) pada peubah prediktor (kriteria seleksi peubah dengan minimal IV sebesar 0,1), akurasi tertinggi yaitu dengan metode *bagging*. Sedangkan, akurasi terendah yaitu metode *logistic regression*. Nilai sensitivitas tertinggi dihasilkan oleh *logistic regression*, sedangkan yang terendah yaitu *random forest*. Nilai spesifisitas tertinggi juga dihasilkan oleh metode *bagging*, sedangkan terendah dihasilkan oleh *logistic regression*. Metode klasifikasi yang dipilih yaitu yang seimbang antara akurasi, sensitivitas, dan spesifisitas. Sehingga untuk pemilihan metode yang akan digunakan untuk prediksi dilakukan penghitungan rataan geometrik terhadap akurasi, sensitivitas, dan spesifisitas. Dari tabel 4, rataan geometrik tertinggi yaitu dengan metode *logistic regression*. Sehingga *logistic regression* lebih baik dibanding metode lainnya jika dilakukan *binning* pada peubah prediktor.

Jika rataan geometrik pada tabel 3 dan tabel 4 dibandingkan maka rataan geometrik tertinggi dihasilkan oleh metode *random forest* yang sebelumnya dilakukan imputasi terhadap data hilang. Sehingga metode ini akan digunakan untuk prediksi terhadap data baru.

3.3 Prediksi Terhadap Data Baru

Hasil pemodelan klasifikasi yang paling baik, yaitu *random forest* selanjutnya digunakan untuk memprediksi data debitur baru. Tersedia file excel yang berisi data debitur baru sebanyak 5339. Hasil prediksi terhadap data baru ini dengan menggunakan hasil metode *random forest* yaitu sebagai berikut.

Tabel 5. Prediksi Status Kolektibilitas Debitur Pada Data Baru

	Status Kolektibilitas		Total
	Baik (*good*)	Buruk (*bad*)	
Jumlah	5168	171	5339
Persentase	96.797	3.203	100

Sumber: Penulis (2023)

Random forest menghasilkan prediksi debitur berkategori baik sebanyak 5168 (96.797 persen) dan berkategori buruk sebanyak 171 debitur (3.203 persen).

4. Kesimpulan

Metode yang memberikan akurasi tertinggi untuk data *dummy* status kolektibilitas debitur dengan *imbalance class* (97.48 persen banding 2.52 persen) dan memiliki banyak *missing value* adalah *random forest* (*missing value* diimputasi dengan nilai rataan). Metode ini menghasilkan akurasi sebesar 80.1346 persen, sensitivitas sebesar 59.3023 persen, dan spesifisitas sebesar 80.67227 persen. Aplikasi *random forest* terhadap data baru menghasilkan prediksi debitur berkategori baik sebanyak 5168 (96.797 persen) dan berkategori buruk sebanyak 171 debitur (3.203 persen).

Saran untuk penelitian selanjutnya adalah agar dilakukan perbandingan imputasi *missing value* dengan berbagai metode, seperti metode rasio, metode ragresi, MICE (*Multiple Imputation by Chained Equations*), dan lain-lain.

Referensi

Agresti, A. (2002). *Categorical Data Analysis*. *John Wiley & Sons, Inc*. New Jersey: John Wiley & Sons, Inc.

Arrahimi, A. R., Ihsan, M. K., Kartini, D., Faisal, M. R., & Indriani, F. (2019). Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa. *Jurnal Sains Dan Informatika*, *5*(1), 21–30. https://doi.org/10.34128/jsi.v5i1.171

Astuti, F. D., & Lenti, F. N. (2021). Implementasi SMOTE untuk Mengatasi Imbalance Class pada Klasifikasi Car Evolution Menggunakan K-NN. *JUPITER*, *13*, 89–98.

Aulia, S., Hadiyoso, S., & Ramadan, D. N. (2015). Analisis Perbandingan KNN dengan SVM untuk Klasifikasi Penyakit Diabetes Retinopati berdasarkan Citra Eksudat dan Mikroaneurisma. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, *3*(1), 75. https://doi.org/10.26760/elkomika.v3i1.75

Dinaloni, D., & Putri, I. C. (2018). Pengaruh Keberlanjutan Usaha Dan Force Majeur Terhadap Kredit Bermasalah Pnpmmandiri Pedesaan Di Kecamatan Trowulan Kabupaten Mojokerto. *Jurnal Pendidikan Ekonomi, Kewirausahaan, Bisnis, Dan Manajemen (JPEKBM)*, *2*(1), 45–60.

Hendrawati, T. (2015). Kajian Metode Imputasi dalam Menangani Missing Data. *Prosiding Seminar Nasional Matematika Dan Pendidikan Matematika UMS*, 637–642. Retrieved from http://hdl.handle.net/11617/5804

Juhola, M., & Laurikkala, J. (2013). Missing values: How many can they be to preserve classification reliability? *Artificial Intelligence Review*, *40*(3), 231–245. https://doi.org/10.1007/s10462-011-9282-2

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. *Science*, *30*(1), 25–36. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248&rep=rep1&type=pdf

Mawarsari, U. (2016). Imputasi Missing Data Dengan K-Nearest Neighbor Danalgoritma Genetika. *AdMathEdu*, *6*(1). https://doi.org/10.12928/admathedu.v6i1.4764

N, S. S., & Sudaryanto. (2022). Sintesis Fitur Density Based Feature Selection (DBFS) Dan Adaboots Dengan Xgboost Untuk Meningkatkan Performa Model Prediksi. *Prosiding Seminar Nasional Sains Dan Teknologi*, 305–313.

Nikmatul Kasanah, A., Muladi, & Pujianto, U. (2017). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Resti*, *1*(3), 196–201.

Prasetio, R. T., & Pratiwi. (2015). Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan Kelas Dataset Medis. *Jurnal Informatika*, *II*(2), 395–403. Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/118

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor. *Jurnal ISD*, *3*(1), 44–49.

Wibawa, A. P., & Dkk. (2018). Metode-Metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134.

Wibowo, A. (2015). Analisis Perbandingan Kinerja Metode Klasifikasi Dalam Data Mining. *Jurnal Integrasi* |, *7*(1), 23–30.

Wijaya, J., Soleh, A. M., & Rizki, A. (2018). Penanganan Data Tidak Seimbang pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB, *2*(2), 32–40.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. London: CRC Press.